eman ta zabal zazu

EUSKAL HERRIKO UNIBERTSITATEA

Hizkuntzaren Azterketa eta Prozesamendua doktoretza-programa

Doktoretza-tesia

Ikasketa-adibide urriko Informazio-Erauzketa

Oscar Sainz Jimenez

2024

eman ta zabal zazu

EUSKAL HERRIKO UNIBERTSITATEA

Hizkuntzaren Azterketa eta Prozesamendua doktoretza-programa

Ikasketa-adibide urriko Informazio-Erauzketa

Oscar Sainz Jimenezek Eneko Agirre eta Oier Lopez de Lacalleren zuzendaritzapean eginiko tesi-txostena, Euskal Herriko Unibertsitatean Doktore titulua eskuratzeko aurkeztua.

Donostia, 2024ko Ekaina.

Simplicity does not precede complexity, but follows it.

— Alan Perlis

Eskerrak

Eskerrik asko...

... Oier eta Eneko, 4 urte hauetan zuzendari bezala eskaini didazuen laguntza guztiagatik. Zuengatik jasotako konfiantza eta aholkuengatik ez balitz, tesi hau ez litzateke atera den bezala aterako.

... Itziar, Oier (berriz), Kepa eta Montse ikerkuntzaren mundua erakusteagatik.

... 318. bulegoko kideei. Bertan gertatutako momentu eta jasotako laguntza guztiarengatik. Eskerrik asko bereziki Iker egindako kolaborazioengatik.

... nola ez, Ixa taldeko kideei. Zaila egiten zait taldean dagoen giroa baino hobeagorik imajinatzea.

... Bonan, for your implication and interest in my work during the early stages of my thesis.

... Hinrich, for hosting me at CIS LMU and giving me the chance to collaborate with you. I learned so much from you during my visit.

... a Sole y Angel por estar siempre ahí, por escuchar mis explicaciones aún cuando os resultaban difíciles de entender y por haberme permitido llegar a donde estoy. Gracias tambien al resto de mi familia y amigos.

Lan hau Eusko Jaurlaritzako diru-laguntza bati esker egin da (PRE_2023_2_013).

Abstract

The field of Information Extraction (IE) allows machines to identify and categorize relevant information that appears in the text, a task that is considered challenging even for humans. Recent advances in the field leverage Machine Learning models trained on large amounts of carefully curated data. However, annotating large corpora is laborious and time-consuming, which can become prohibitive in low-resource scenarios.

The goal of this thesis is to explore and develop IE methods that work on low-resource scenarios. Particularly, exploiting the generalization capabilities of Language Models (LM), and how they allow knowledge transfer between high-resource and low-resource scenarios. The thesis is comprised of two main parts. Briefly, in the first part, we have developed a Zero- and Few-shot Information Extractor using encoder-only Language Models. For the second part, we have jumped to Large Language Models and addressed some limitations of our previous approach.

More in detail, in order to develop a Zero- and Few-shot Information Extractor it is necessary to bypass the need for large amounts of schema-dependant annotated data. Recent advances in the field have shown that many tasks can be reformulated as natural language or as other high-resource tasks, and be solved by pre-trained Language Models. In this thesis, we propose the use of Textual Entailment as an intermediate task for Information Extraction. By reformulating the tasks as Textual Entailment, the model is no longer tied to a specific schema, allowing it to generalize and work as a zero-shot information extractor out of the box. We have shown that this approach can achieve zero-shot results similar to supervised systems from a few years ago. Moreover, the model can be further trained with examples of the task, achieving results close to the supervised state-of-the-art. Removing the schema dependency allows the model to be trained on multiple schemas —different datasets— at the same time. We have shown that

for datasets from similar tasks, a model trained on one schema can transfer the knowledge to another schema, improving the zero-shot performance and further reducing the gap with supervised systems.

Reformulating Information Extraction tasks as Textual Entailment requires some manual work, the instances of the task need to be converted to premise-hypothesis pairs. To that end, manually created templates are used to generate the hypotheses, which we call verbalizations. As part of this thesis, we have shown that the effort of creating verbalizations is significantly lower than the effort of creating annotations for the same task, yielding better performance for the same amount of work. We have also shown that different experts can make the verbalizations —with different styles— and still perform similarly. Based on these results, we have proposed a new workflow —verbalize while defining— which we compare with traditional *define, annotate, and train* workflows. This workflow allows novice users to model complex IE schemas with strong zero-shot performance that later can be curated by expert annotators with much less effort. We have developed a practical demonstration of the proposed workflow.

For the last part of this thesis, we analyzed and addressed the limitations of the Textual Entailment approach. We leveraged the progress carried out by decoder-only LLMs and implemented GoLLIE: an LLM capable of following annotation guidelines to perform IE annotations. Different from the Textual Entailment approach, GoLLIE leverages detailed guidelines —thanks to the longer context window— instead of simple verbalizations, allowing the model to follow more fine-grained instructions. Additionally, we provided an error analysis and future research directions for the field.

Note for non-Basque-speaking readers: This dissertation is structured as a collection of articles. The introductory chapter is in Basque, whereas the conclusions and articles are in English. Non-Basque-speaking readers are recommended to read the Conclusions chapter to get an overview of the main contributions made to this thesis, followed by the papers in the appendix in their recommended reading order.

Laburpena

Informazio-erauzketaren arloak makina bati testuan agertzen den informazioa identifikatzea eta sailkatzea nola irakatsi ikertzen du. Ataza hau, gizakiontzat ere erreza ez dena, azken urteetan ikasketa-automatikoan egindako aurrerapenek sustatu dute, datu-anotatuak erabiliz ereduak entrenatuz. Hala ere, corpora handiak anotatzea lan neketsu eta garestia da, batez ere baliabide urriko inguruneetan.

Tesi honen helburua baliabide urriko inguruneetan IE metodoak aztertzea eta garatzea da. Zehazki, hizkuntza-ereduen orokortze gaitasunak erabiltzea, batez ere baliabide handiko iturrietatik ikasitakoa baliabide urriko inguruneetara transferitzeko gaitasuna. Tesia bi zati nagusietan banatzen da. Laburki, lehenengo zatian, hizkuntza-eredu kodetzaileak erabiliz ikasketa-adibiderik gabeko edo urriko sistema bat garatu da informazio-erauzketa gauzatzeko gai dena. Bigarren zatian, hizkuntza-eredu handiagoetara salto egin da eta aurreko metodoaren zenbait mugarri aztertu dira.

Zehatzago, adibiderik gabeko edo urriko informazio-erauzketa sistema bat garatzeko beharrezkoa da eskemen menpe dauden datu-anotatu kopuru handien beharra gainditzea. Hizkuntzaren prozesamenduan azkenaldian egon diren aurrerapenei esker ikusi da ataza asko birformulatu daitezkeela testu-sorkuntza edo bestelako baliabide handiko ataza batzuetara, eta, honi esker hizkuntza-ereduen bitartez ebatzi ahal izatea. Tesi honetan testuzko inferentzia erabiltzea proposatzen da informazio-erauzketa ebazteko erdibideko ataza bezala. Testuzko inferentzia bezala birplanteatuz entrenatutako eredua ez da gehiago eskema bati lotuta egongo. Horrela, ataza edo eskema berrietara orokortzea eta adibiderik gabe informazio-erauzketa gauzatzea lortuz. Tesi honetan erakutsi da hurbilpen honek duela urte batzuetako sistema gainbegiratuen emaitzak berdintzea lortzen duela ikasketa adibiderik erabili gabe. Ez hori bakarrik, proposatutako hurbilpena adibide gutxi batzuekin entrenatuz gero, gaur egungo artearen egoera berdintzen duela erakutsi da. Anotazio-eskemarekiko menpekotasuna ezabatzean erakutsi da eredua hainbat eskemetatik —datu-multzoetatik— ikas dezakeela aldi berean. Ataza antzekoetako datu-multzoentzat, eskema batean entrenatutako eredu batek beste eskemara ezagutza transferitu dezake, adibiderik gabeko eta sistema gainbegiratuen arteko emaitzen desberdintasuna are gehiago txikituz.

Informazio-erauzketako atazak testuzko inferentzia atazara birformulatzeko eskuzko lana behar da, atazaren adibideak premisa-hipotesi pareetara bihurtu behar direlako. Helburu horretarako, eskuz sorturiko txantiloiak erabiltzen dira lehendabizi automatikoki hipotesiak sortzeko. Txantiloi hauek *testuzko adierazpen* deitu dira. Tesi honetan erakutsi da testuzko adierazpen hauek sortzeko esfortzua esanguratsuki txikiagoa dela informazio-erauzketako adibideak anotatzea baino, errendimendu askoz hobea lortuz esfortzu berdinerako. Horretaz gain, erakutsi da jakintza arlo desberdinetako adituek sortutako testuzko adierazpenekin —estilo desberdina dutenak— emaitza antzekoak lortzen direla. Emaitza hauetan oinarrituta, lan fluxu berri bat —*verbalize-while-defining* deiturikoa— proposatu da eta definitu, anotatu eta entrenatu lan-fluxu tradizionalarekin konparatu da. Lan-fluxu berri honek, erabiltzaile berri bati ahalbidetzen dio informazio-erauzketa eskema konplexuak garatzea errendimendu altu batekin, horrela gaiko adituek ereduak sortutako anotazioak esfortzu gutxiagorekin zuzendu ditzaten. Proposatutako lan-fluxua praktikan erakusteko prototipo bat garatu da.

Tesiaren azkeneko zatian, testuzko inferentzian oinarritutako hurbilpenak dituen mugak aztertu eta batzuk konpondu dira. Hizkuntza-eredu deskodetzaileek ekarri dituzten abantailak baliatu dira eta GoLLIE garatu da: informazio-erauzketako anotazio gidalerroak jarraitzeko gai den hizkuntza-eredu handia. Testuzko inferentzian oinarritutako hurbilpenak ez bezala, GoLLIE-k anotazio gidalerro detailatuak —testuinguru luzera handiagoari esker— erabiltzen ditu testuzko adierazpen sinple batzuen ordez, ereduari detaile handiagoko instrukzioak jarraitzea ahalbidetuz. Gainera, errore-analisia egin eta etorkizuneko ikerketa-lerroak proposatu dira.

Gaien aurkibidea

Irudien zerrenda

1. KAPITULUA

Sarrera

Tesi-txosten hau artikulu-bilduma bezala dago antolatuta. Atal honetan tesiaren aurkezpena egiten da, hurrengo egitura jarraituz: 1.1 atalean landutako gaiaren motibazioa aurkezten da. 1.2 atalean tesi honen ikerketa-lerro eta helburuak deskribatzen dira, gero, 1.3 atalean egindako kontribuzioak zerrendatzeko. 1.4 eta 1.5 ataletan tesian erabilitako kontzeptuen oinarriak eta erlazionatutako lanak aurkezten dira hurrenez-hurren. Atal honen ondoren, 2 kapituluan tesian ateratako ondorioak deskribatu eta etorkizuneko lan nagusiak eztabaidatzen dira. Azkenik, A apendizean tesi hau osatzen duten artikuluak aurkezten dira.

1.1 Motibazioa

Informazio-erauzketa (IE) testu huts batetik garrantzizko informazioa identifikatu eta modu egituratu batean erauztean datza. Makinek —hizkuntza-eredu handien aurretik behintzat— ez dira kapaz gizakiok erabiltzen dugun hizkuntza zuzenean ulertzeko. Hori dela eta, beharrezkoa izaten da testu bat aurre-prozesatzea bertako informazioa erabili nahi denean. Informazio-erauzketa hizkuntzaren prozesamenduaren ikerkuntza-arloan historikoki garrantzi handia izan duen gai bat da bereziki horregatik.

Gizakiok ohituta gaude, testu bat irakurtzen dugun heinean garrantzizko datuak —entitateak eta gertaerak— identifikatzen. Norberak irakurtzen duen informazioaren datu-egitura bat sortzen du. Helburuaren arabera eta erabilgarritasun ikuspuntu batetik ordea, garrantzitsua izan daiteke informazio hori aurretik definitutako eskema edo zehaztapen bat jarraitzea. Informazio-erauzketan hain zuzen hori egiten da: erauzi nahi den informazioaren eskema anotazio gidalerroetan definitzen da. Aditu batentzat nahikoa da gidalerro hauek —zehaztapen eta adibide gutxi batzuk— irakurtzea ataza modu zuzen batean ebazteko gai izateko. Gaur egungo[1] hizkuntzaren prozesamenduko tresnek ordea ez dute ahalmen hori.

Informazio-erauzketako artearen egoera anotatutako datuak erabiltzen dituzte ikasketa-automatikoko ereduak entrenatzeko. Ereduek, gidalerroetan definituta dauden zehaztasun eta salbuespenak datuetatik ikasten dituzte. Hori horrela, datuak dira errendimendu ona lortzeko faktore garrantzitsuena. Datu hauen sorkuntzak eskuzko lana eskatzen du, eskuzko lan handia hain zuzen. Lehenik, ataza definitu eta gidalerroak sortu behar dira. Gero, anotatzaileak izango diren adituak trebatu eta hauek dituzten desadostasunak ebatzi behar dira. Azkenik, testu kopuru handi bat anotatu beharra dago (Walker et al., 2006). Prozesu hau ez da merkea, eta praktikan oso datu-multzo gutxi daude kalitatezkoak eta handiak direnak aldi berean. Honek eragin du informazio-erauzketan datu-sorkuntzako tekniken garapen eta ikerkuntza bultzatzea (Mintz et al., 2009).

Tesi proiektu hau planteatu zen garaian, paradigma aldaketa baten hasieran zegoen hizkuntzaren prozesamenduko ikerkuntza-arloa. Hizkuntza-ereduek, bereziki hizkuntza-eredu handiek, protagonistak izan diren paradigma aldaketa bat izan da (Min et al., 2023). Eredu hauek, testu bat emanda hurrengo hitza asmatzeko entrenatzen dira. Ataza hau erreza dirudien arren, bere zailtasuna du. Alegia, helburu honi esker hizkuntza ereduek testuinguruan beste ataza baten adibide batzuk emanda gai dira ataza ebazten ikasteko (Brown et al., 2020). Aurkikuntza

[1] Tesia planteatu zen garaian behintzat.

hau, hizkuntzaren prozesamenduko ikerkuntza-arlo osoa birplanteatzera eraman du, baita informazio-erauzketaren arloa ere.

Tesi honetan, aurreko paradigma aldaketak bultzatuta informazio-erauzketak duen datu beharrari aurre egiteko tekniken ikerkuntza eta garapena egin da. Zehazki, datu anotaturik behar ez dituzten edo gutxi behar dituzten sistemen garapena ikertu da. Ikerketa-lerro hau justifikatzeko bi arrazoi nagusi nabarmendu daitezke:

Interes zientifikoa. Informazio-erauzketaren ataza zuzen ebazteko, testuaren egituraren eta esanahiaren ulermena beharrezkoa da. Datuetan oinarritutako sistemek, datuetan ageri diren patroi estatistikoak baliatzen dituzte ataza zuzen ebazteko. Atazerako bereziki entrenatuak izan ez diren ereduek ordea, ataza ebazteko behar duten ahalmena beste nonbaiten ikasi eta bertan aplikatzeko gaitasuna eduki behar dute. Hau da, dagoeneko zaila den ataza batera daturik gabe orokortzeko gaitasuna erakutsi behar dute. Zentzu honetan atazen arteko ezagutza-transferentzia ikertzen duen ikerketa-lerro bat da. Hori horrela, ikerketa-lerro honek ahalbidetzen du hizkuntza-ereduen gaitasunen ulermenean sakontzeko.

Praktikotasuna. Tesi honen motibazio nagusia alde praktikotik dator batez ere. Izan ere, hain garestiak diren datu anotatuen beharra gutxitzea edo guztiz ezabatzea ekarpen esanguratsua da. Ez hori bakarrik, dagoeneko existitzen diren datuak ataza edo eskema berrietan berrerabiltzeko aukera eskuzko lan asko murriztu dezake. Azkenik, anotatutako daturik behar ez dituzten sistemek eskatu ahalako (*on-demand* ingelesez) informazio-erauzketa gauzatzea ahalbidetuko lukete. Oro har, ikerketa-lerro hau informazio-erauzketaren arloak jasaten duen arazo nagusietako bati aurre egiten dio.

1.2 Helburuak eta ikerketa-lerroak

Tesi honen helburua informazio-erauzketa arloan ikasketarako daturik gabe edo gutxirekin ataza gauzatzeko gai diren sistemak garatzea izan da. Tesia hasteko garaian adibide urriko hurbilpenak testu sailkapeneko atazetara mugatuta zeuden. Informazio-erauzketaren kasuan ordea, zehaztapen konplexuko atazak izanda existitzen ziren hurbilpen horiek ez zuten arrakasta handirik izan. Helburua beraz, existitzen diren hurbilpenak informazio-erauzketara nola egokitu, edo beharrezkoa bada hurbilpen berriak garatzea izan da. Tesi honetan informazio-erauzketaren lau ataza nagusiak aztertu dira: izendun entitateen erauzketa, erlazio-erauzketa, gertaera-erauzketa eta gertaeren argumentu-erauzketa.

Tesiaren garapenean zehar hizkuntzaren prozesamenduko arloa asko eraldatu da, abiadura handi batean. Tesi hau GPT-3 —175 mila milioi parametrodun hizkuntza-eredu bat— argitaratu eta gutxira hasi zen, oraindik ere ehundaka milioi parametrodun hizkuntza-ereduak erabilienak izanda. Hori dela eta, tesi honen hasieran kodetzaile arkitekturan oinarritutako hizkuntza-ereduekin egin da lan. Tesiaren azken zatian aurretik ikasitakoa gaur egun artearen egoera diren hizkuntza-eredu deskodetzaileetara aplikatu da. Zehazki, tesi honetan garatu diren ikerketa-lerroak hurrengoak izan dira:

[L1] **Ikasketa-datu gabe edo gutxirekin informazio-erauzketa gauzatzeko gai den sistema baten garapena**, hizkuntza-eredu kodetzaileak erabiliz. Ikerketa-lerro honetan artearen egoera aztertu eta daturik behar ez duen informazio-erauzketako eredu bat proposatu da. Lan hori arloko berri honen lehendabizikoetariko bat izan da. Ikerketa lerro hau hiru azpilanetan banandu da: daturik behar ez duen hurbilpenaren garapena, datu gutxirekin hurbilpena hobetzeko tekniken azterketa eta hurbilpenaren aplikagarritasuna praktikara eramatea.

[L1.1] **Ikasketa-daturik behar ez duen informazio-erauzketa sistema baten garapena.** Artearen egoerako hurbilpenek hizkuntza-eredu bati parametro berri batzuk gehitu eta datu gainbegiratuak erabiliz atazara egokitzen dute. Estrategia hau ordea, daturik ez daudenean ausazko iragarpenak egiten ditu. Lan lerro honetan tesi honen muina den hurbilpena garatu da, zehazki, testuzko inferentzian oinarritutako informazio-erauzketarako hurbilpena. Aurreko sistemak ez bezala, eredua atazara zuzenean egokitu beharrean, helburu-ataza bera da egokitzen dena. Alegia, burutu nahi den ataza ereduak dagoeneko dakien ataza batera —pibot atazara— bihurtzen da. Horri esker, eredua zuzenean aplikatu daiteke ataza ebazteko.

[L1.2] **Adibide-urriko ikasketa eta anotazio-eskemen arteko transferentzia.** Lan lerro honetan testuzko inferentzian oinarritutako eredua datu gutxi batzuekin fintzeko aukerak aztertu dira. Horretarako, helburu atazaren adibideak pibot atazara bihurtzeko estrategia automatiko bat proposatu eta ebaluatu da. Ikasketa pibot atazean gauzatzeak helburu atazarekiko menpekotasuna apurtzen du, ataza —edo eskema— desberdinetatik ikasteko aukera emanez. Hori horrela, eskemen arteko ezagutza-transferentzia aztertu da pibot-ataza bidezko ikasketaz baliatuz.

[L1.3] **Hurbilpenaren aplikazio praktikoa** eta demo baten garapena. Lan lerro honetan garatutako hurbilpenak dauzkan abantailak ikuspuntu praktiko batetik nola aprobetxa daitezkeen aztertu da. Abantaila nagusia eskema berri batentzat ikasketa daturik gabe ere funtzionatzeko gaitasuna da. Hori horrela, informazioa erauzteko lan-fluxu berri bat proposatu da, eta bere erabilgarritasuna erakusteko demo bat garatu da.

[L2] **Hizkuntza-eredu handiak eta adibiderik gabeko informazio-erauzketa.** Hizkuntza-eredu handiek —mila milioi parametro baino gehiago dituzten hizkuntza-ereduek— aurrerapen esanguratsuak erakutsi dituzte, batez ere adibiderik gabeko ikasketan. Ikerketa-lerro honetan, giza anotatzaileek erabiltzen dituzten gidalerroak jarraitzeko hizkuntza-eredu handien gaitasuna aztertu da. Aurkitu diren oztopoen artean, ikusi da hizkuntza-ereduek —gizakion antzera— kontzeptuen aurreiritzi sendo bat dutenean gidalerroak kontutan hartzeko zailtasunak dituztela, anotazio akatsak sortuz. Adibidez, pertsona kontzeptua ezagutzen dutenez, emandako definizioari kasu ez egitea erabakitzen dute. Ikerketa-lerro honen barruan arazo honi aurre egiteko teknikak garatu dira.

1.3 Egindako kontribuzioak

Atal honetan tesian egindako kontribuzioak argitaratutako artikuluka aurkezten dira. 1.3.1. atalean tesia osatzen duten artikuluak aurkezten dira, irakurketa-orden gomendatuan. 1.3.2. atalean berriz tesian zehar argitaratutako baina tesiaren parte ez diren artikuluak zerrendatzen dira, orden kronologikoan.

1.3.1 Tesia osatzen duten artikuluak

Atal honetan tesia osatzen duten artikuluak aurkezten dira. Artikulu bakoitza bere erreferentzia, laburpena eta dagokion ikerketa-lerro eta tesiaren testuinguru orokorrean kokatzen duen azalpenarekin batera aurkezten da. Artikulu osoak A eranskinean aurki daitezke.

[**A.1**] Sainz et al. (EMNLP 2021)

Sainz O., Lopez de Lacalle O., Labaka G., Barrena A., and Agirre E. Label Verbalization and Entailment for Effective Zero and Few-Shot Relation Extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A.1. artikulua L1.1 eta L1.2. ikerketa-lerrotan kokatzen da, hau da, adibiderik gabeko informazio-erauzketa sistema baten garapenean eta adibide-urriko ikasketan. Artikulua hau tesiaren lehenengoa izan da eta datozen artikuluen oinarriak ezarri ditu. Lan honen motibazioa tesiaren motibazio nagusietatik dator: informazio-erauzketako datuak anotatzeko oso garestiak dira eta gaur egungo sistemek adibide asko behar dituzte behar bezala funtzionatzeko. Ondorioz, datu anotatu gutxiago darabilten sistemen garapena motibatzen du. Artikulu honetan aipatutako motibazioa egia dela erakusteko artearen egoerako ereduak ebaluatu dira datu-kopuru murriztuko erregimen batean. Horretarako, TACRED (Zhang et al., 2017) erlazio-erauzketarako datu-multzoa erabili da. Analisiak erakutsi du metodo hauek azkar kaltetzen direla datuak murriztuz gero.

Hori horrela, testuzko inferentzian oinarritutako erlazio-erauzketarako hurbilpena proposatu da. Hurbilpenaren ideia dagoeneko landua izan da tesi honen aurreko artikulu batean (Sainz and Rigau, 2021), A.1. artikuluan ordea metodo hori egokitu da erlazio-erauzketara aplikatzeko. Hurbilpenaren kalitatea neurtzeko hiru eszenatoki definitu dira: ikasketarako daturik ez egotea, ikasketarako datu gutxi batzuk egotea, eta ikasketarako datu asko egotea. Lehenengo kasurako testuzko inferentziako eredua zuzenean aplikatu da ebaluazio-multzoaren gainean. Bigarren eszenatokirako, eredua adibide gutxi batzuekin entrenatua da, datu-multzo originalaren %1, %5 eta %10-arekin hain zuzen. Ebaluaketa honen emaitzak, artearen egoerako sistemen emaitzekin alderatu dira, alde handi batez emaitzak hobetuz. Nabarmentzekoa da ere artearen egoerako sistemek datu gutxi batzuk erabili arren ere ez direla gai proposatutako sistemak adibiderik gabe lortzen dituen emaitzetara iristeko. Azkenik, datu-multzoaren datu guztiekin entrenatu dira sistema guztiak. Kasu honetan proposatutako sistema lehiakorra da ere, baina artearen egoerako sistemekin konparatuta aldea ez da hain handia.

Proposatutako hurbilpena adibiderik gabe emaitza onak lortzeko kapaz bada ere, inferentzia-abiadura aldetik ez da oso azkarra. Alegia, ebaluazio adibide bakoitzeko hitzezko adierazpen haina inferentzia egin behar dituelako da (ikusi A.1. artikulua). Hurbilpenaren gaitasunak mantenduz artearen egoerako sistemen abiadura lortzen duen sistema bat garatzeko ezagutza-destilazioa —eredu arrunt bat entrenatu da testuzko inferentzia sistemak egindako anotazioen gainean— aplikatu da. Lortutako emaitzek erakutsi dute bi sistemek, testuzko inferentzian oinarritutako hurbilpena eta destilazio bidez entrenatutako eredua, parekoak direla gaitasun aldetik bigarrena inferentzia garaian azkarragoa izanda.

[**A.2**] Sainz et al. (NAACL-Findings 2022)

Sainz O., Gonzalez-Dios I., Lopez de Lacalle O., Min B., and Agirre E. Textual Entailment for Event Argument Extraction: Zero- and Few-Shot with Multi-Source Learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2439–2455, Seattle, United States. Association for Computational Linguistics.

A.2. artikulua L1.2. ikerketa-lerroan kokatzen da, adibide-urriko ikasketan eta anotazio-eskemen arteko transferentzian. A.1. artikuluan lortutako emaitzak ikusita beste atazetara aplikatzea erabaki da. Kasu honetan, gertaeren argumentu-erauzketara.

Gertaeren argumentu-erauzketa erlazio-erauzketa baino konplexuagoa da hitzezko adierazpenak sortzeko garaian. Hori horrela, testuzko inferentzian oinarritutako hurbilpena beste ataza konplexuagoetan funtzionatzeko gai den aztertu nahi izan da. Kasu honetan, bi dira erabili diren ebaluazio datu-multzoak: ACE05 (Walker et al., 2006) eta WikiEvents (Li et al., 2021). Egindako ebaluazioak erakutsi du metodoaren eraginkortasuna.

Testuzko inferentzian oinarritutako hurbilpenak ezaugarri berezi bat du: ez dago anotazio eskema bati lotuta. Ezaugarri horrek ahalbidetzen du ikasketa-daturik gabe ere aplikagarri izatea, eta aldi berean, eskema edo ataza bat baino gehiagotik ikasi ahal izatea. Beraz, bi galdera nagusi aztertu dira A.2. artikuluan: *onuragarria al da testuzko inferentzia datu-multzo bat baino gehiagotik ikastea? Posiblea al da eskema batetik bestera ezagutza-transferentzia gauzatzea eskuzko mapaketarik erabili gabe?* Lehenengo galdera aztertzeko MNLI (Williams et al., 2018) datu-multzoan soilik entrenatutako eredu bat SNLI (Bowman et al., 2015), MNLI, Fever-NLI (Thorne et al., 2018) eta ANLI (Nie et al., 2020) datu-multzoetan entrenatutako eredu batekin alderatu da. Emaitzek erakutsi dute kasu guztietan onuragarria dela ahal den eta datu-multzo gehien erabiltzea. Bigarren galderarako, helburu eskema baten gainean aplikatu aurretik bitarteko entrenamendu bat gauzatu da, adibidez: ACE –> WikiEvents —ACE datu-multzoan entrenatu eta WikiEvents datu-multzoan ebaluatu— edo WikiEvents –> ACE. Egindako ebaluazioak frogatu du testuzko inferentziaren hurbilpenaren bitartez posiblea dela eskema batetik bestera ezagutza transferitzea. Ondorioz, emaitzak nabarmenki hobetuz, bereziki ikasketa-daturik gabeko eszenatokietan. Bestalde, ataza desberdinen —erlazio-erauzketa eta gertaeren argumentu-erauzketaren— arteko transferentzia ere aztertu da, horretarako A.1. artikuluan erabilitako TACRED datu-multzoa erabili da. Kasu honetan ordea, lortutako hobekuntzak ez dira esanguratsuak izan.

Azkenik, beharrezkoa den eskuzko lanaren eragina aztertu da. Hurbilpena aplikatu ahal izateko erlazio edo argumentuen hitzezko adierazpenak sortu behar dira. Hauek ordea eskuz sortutako txantiloien bitartez sortzen dira eta estrategia desberdinak jarrai daitezke. Estrategia desberdinen eragina neurtzeko bi adituen txantiloiak alderatu dira. Txantiloiak sortzeko garaian, adituek 15 minutu izan dituzte argumentu mota bakoitzeko nahi haina txantiloi sortzeko atazaren gidalerroak eskura izanda. Jarraitu dituzten estrategiak nabarmenki desberdinak izan diren arren, ereduak lortzen dituen emaitzak parekoak izan dira. Bestetik, hitzezko adierazpenak sortzeko eta anotazioak egiteko behar den esfortzua konparatu da, hauek sortzeko beharrezkoa den denbora eta ereduak lortzen dituen emaitzen menpe. Azterketa hau egiteko ACE datu-multzoko adibide batzuk eskuz anotatu dira denboraren estimazio bat edukitzeko. Ondorio nagusia izan da merezi duela txantiloiak sortzeari lehentasuna ematea hasieran, eta, gainontzeko denbora ikasketarako adibideak anotatzen erabiltzea komeni dela.

[**A.3**] Sainz et al. (NAACL 2022)

Sainz O., Qiu H., Lopez de Lacalle O., Agirre E., and Min B. ZS4IE: A toolkit for Zero-Shot Information Extraction with simple Verbalizations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 27–38, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

A.3. artikulua L1.3 ikerketa-lerroan kokatzen da, hurbilpenaren aplikazio praktikoan eta demo baten garapenean. Tesi honen kontribuzio nagusia informazioa erauzteko metodo berri bat garatzea izan da, ikasketa-daturik behar ez dituena. Izan ere, A.2. artikuluan ikusi da hitzezko adierazpenak sortzeko txantiloiak sortzea adibideak zuzenean anotatzea baino eraginkorragoa dela. Artikulu honetan, ideia horretan oinarrituta lan-fluxu berri bat inplementatzen duen demo bat garatu da.

Definitutako lan-fluxu berria hurrengo pausutan oinarritzen da: (1) interesezko kontzeptua —entitatea, erlazioa, gertaera edo argumentua— definitzen da txantiloi batzuk sortuz, (2) adibide batzuk gehitzen dira eta sistema ebaluatzen da, (3) txantiloi bakoitzak izan duen eragina —positiboa edo negatiboa— aztertu eta beharrezkoak diren aldaketak aplikatzen dira; (4) prozesua iteratu beste kontzeptu batentzat. Aurretik existitzen den lan-fluxuarekin alderatuta; hau da, kontzeptuak zehatz-mehatz definitu, testuak anotatu eta ikasketa-automatikoko eredu bat entrenatzearekin konparatuz bi dira lan-fluxu berriaren onurak. Alde batetik, prozesua iteratiboa izanda anotatu nahi den eskeman zerbait aldatu —edo gehitu— nahi izanez gero prozesua erreza eta azkarra da, aldiz lan-fluxu tradizionalean anotazioa eta entrenamendu faseak osorik errepikatu beharko lirateke. Bestetik, txantiloiak sortzea sormenezko ataza bat izanda atseginagoa da datuak anotatzea baino.

Garatutako sistema informazio-erauzketaren lau ataza nagusietan ebaluatu da. Hori horrela, lehendik aipatutako datu-multzoetaz aparte entitate izendunen erauzketarako CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) datu-multzoa erabili da. Gertaera-erauzketa ataza ebaluatzeko berriz, dagoeneko aurretik erabilia izan den ACE05 datu-multzoa erabili da.

[**A.4**] Sainz et al. (ICLR 2024)

Sainz O., García-Ferrero I., Agerri R., Lopez de Lacalle O., Rigau G., and Agirre E. GoLLIE: Annotation guidelines improve zero-shot information-extraction. *The Twelfth International Conference on Learning Representations*, Vienna, Austria, May 2024.

A.4. artikulua L2. ikerketa-lerroan kokatzen da, hizkuntza-eredu handiak eta adibiderik gabeko informazio-erauzketan. Artikulu hau eta aurrekoen artean hizkuntzaren prozesamenduko arlo osoari eragin dion belaunaldi aldaketa bat gertatu da: iturri irekiko hizkuntza-eredu handien agerpena. Aldaketa hau tesi honetan ere eragina izan du.

A.1. artikuluan proposatutako sistema, A.2. eta A.3. artikuluetan erabili dena, diseinu aldetik hainbat muga dauzka. Adibidez, badaude kasu asko non testuzko inferentziarako gaitasunekin soilik ezin direnak ondorioztatu. Esaterako "Peterrek sagar bat jan zuen" esalditik ezin da ondorioztatu "Peter pertsona bat da" hipotesia egia dela. Zehaztasunarekin zerikusia duten mugak ere aurki daitezke, besteak-beste hitzezko-adierazpen bakunetan erabil daitekeen zehaztasun maila txikia. Hau da, eskemari buruz adieraz daitekeen informazio kopurua mugatua da. Horrek eragin handia izan dezake batez ere mota bera duten datu-multzo edo eskema desberdinen arteko desberdintasunak adierazteko garaian. Azkenik, muga tekniko batzuk ere baditu, hain zuzen sailkapena egiteko hautagaien beharra edota beharrezkoak diren inferentzia kopurua.

Arazo hauei aurre egiteko, A.4. artikuluan anotazio gidalerroak jarraitzeko gai den hizkuntza-eredu handietan oinarritutako hurbilpen bat proposatu da. GoLLIE, ***G**uideline following **L**arge **L**anguage model for **I**nformation **E**xtraction* ingelesezko sigletatik, giza anotatzaileek erabiltzen dituzten gidalerroak baliatzen ditu informazio-erauzketako atazak ebazteko. Hizkuntza-eredu handiek erabiltzen duten testuinguru luzera handiak ahalbidetzen du detaile handiagoko azalpenak hizkuntza-ereduari ematea, hala ere, hori bakarrik ez da nahikoa hizkuntza-ereduak gidalerroak jarrai ditzan. Proposatutako hurbilpenean, hainbat estrategia desberdin baliatzen dira ereduari gidalerroak jarraitzen ikasarazteko. 10 datu-multzoen gainean egindako ebaluazioak erakutsi du proposatutako hurbilpenaren eraginkortasuna ikasketa-adibide gabeko kasuetan. Kasu gainbegiratuan ere, beste 9 datu-multzoetan, artearen egoerako emaitzak lortu dira. Bukatzeko, errore analisi bat egin da etorkizuneko ikerketa-lerro posibleak aztertzeko. Bertan, ereduak dituen zailtasun espezifikoetaz aparte datu-multzo batzuek dituzten diseinu arazo batzuk ere aurkitu dira.

1.3.2 Tesia osatzen ez duten artikuluak

Atal honetan tesia osatzen ez duten artikuluak zerrendatzen dira. Artikulu bakoitza bere erreferentzia eta laburpenarekin batera aurkezten da. Artikulu hauek ez dira A. apendizera gehitu.

Sainz and Rigau (GWC 2021)

Sainz O. and Rigau G. Ask2Transformers: Zero-shot domain labelling with pretrained language models. In Vossen P. and Fellbaum C., editors, *Proceedings of the 11th Global Wordnet Conference*, 44–52, University of South Africa (UNISA), January 2021. Global Wordnet Association.

Laburpena: Artikulu honetan gainbegiraketarik gabe WordNet-eko adierei domeinu bat esleitzeko gai den aurre-entrenaturiko hizkuntza-ereduetan oinarritutako sistema bat aurkezten da. Gainera, sistema ez dago mugatuta aurre-definituriko domeinu etiketa multzo batera. Hurbilpena hizkuntza-eredu aurre-entrenatuetan kodetuta dagoen ezagutzaz baliatzen da ataza birformulatuz domeinu etiketak esleitzeko. Proposatutako sistema ebaluazioan erabilitako ingelesezko datumultzoan artearen egoera berri bat ezartzen du adibiderik gabeko inferentzian.

Sainz et al. (GWC 2023)

Sainz O., de Lacalle O.L., Agirre E., and Rigau G. What do language models know about word senses? zero-shot WSD with language models and domain inventories. In Rigau G., Bond F., and Rademaker A., editors, *Proceedings of the 12th Global Wordnet Conference*, 331–342, University of the Basque Country, Donostia - San Sebastian, Basque Country, January 2023b. Global Wordnet Association.

Laburpena: Gaur egun hizkuntzaren prozesamenduko sistema gehienen muinean daude hizkuntza-ereduak. Hitz-bektoreak ez bezala, hizkuntza-ereduek testuingurudun adierazpenak erabiltzen dituzte, ezinbesteko ezaugarri bat adierak desanbiguatzeko. Artikulu honetan, inferentzia garaian hizkuntza-ereduek adierak desanbiguatzeko zenbateraino gai diren esploratzea dugu helburu. Analisia burutzeko BERT edo RoBERTa moduko hizkuntza-ereduak testuz baldintzatu ditugu hitz-adieren desanbiguazio ataza egikaritzeko. Horretarako, adiera eta domeinuen arteko erlazioak erabili ditugu, eta, ataza testuzko inferentzia balitz bezala planteatu dugu, non hipotesi bakoitzak adiera bakoitzaren domeinuari egiten dio erreferentzia. Emaitzek erakusten dute hurbilpen hau hain zuzen efektiboa dela, emaitza gainbegiratuetara hurbilduz.

Min et al. (ACM Computing Surveys 2023)

Min B., Ross H., Sulem E., Veyseh A.P.B., Nguyen T.H., **Sainz O.**, Agirre E., Heintz I., and Roth D. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2), sep 2023. ISSN 0360-0300.

Laburpena: BERT eta GPT bezalako hizkuntza-eredu aurre-entrenatuek hizkuntzaren prozesamenduko arloa guztiz eraldatu dute. Ataza askotarako, hizkuntza-eredu handietan oinarritutako hurbilpenek artearen egoera ezarri dute. Ataza orokor batetik hizkuntzaren adierazpen orokorrak behin ikastea da gakoa gero beste ataza desberdinetara hedatu ahal izateko. Ataza orokor bezala hizkuntza-eredu ikasketa erabiltzen da, auto-gainbegiraketarako testuzko datu asko dituena ikasketa luze bat egiteko. Artikulu honetan hizkuntza-ereduen arkitekturen fundamentuzko kontzeptuak eta hizkuntza-ereduak erabiltzen dituzten teknikak aurkezten dira ikuspuntu orokor batetik. Aurre-entrenatu eta findu, testuzko baldintzazko inferentzia (*prompting* ingelesez) eta testu sorkuntzarako hurbilpenak jaso dira. Horrez gain, hizkuntza-ereduek dituzten mugak eta etorkizuneko ikerketa-lerroak eztabaidatzen dira.

García-Ferrero et al. (SemEval 2023)

García-Ferrero I., Campos J.A., **Sainz O.**, Salaberria A., and Roth D. IXA/cog-comp at SemEval-2023 task 2: Context-enriched multilingual named entity recognition using knowledge bases. In Ojha A.K., Doğruöz A.S., Da San Martino G., Tayyar Madabushi H., Kumar R., and Sartori E., editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 1335–1346, Toronto, Canada, July 2023. Association for Computational Linguistics.

Laburpena: Izendun entitateen erauzketa hizkuntzaren prozesamenduko oinarrizko ataza bat da non hizkuntza-eredu aurre-entrenatuek emaitza aipagarriak lortu dituzten. CoNLL 2003 bezalako ebaluazio-multzo estandarrek ordea, ez dituzte zabalduta dauden entitate-erauzketa sistemen erronkak islatzen; entitate berri eta konplexuak modu oso zehatz batean sailkatzea adibidez. Artikulu honetan, hiru pausuko izendun entitateen erauzketarako hurbilpen teilakatu bat aurkezten da: lehendabizi, sarrera esaldian entitateak izateko hautagaiak topatzen dira; bigarren, hautagai bakoitza ezagutza-base bateko sarrera batekin lotzen da; azkenik, hautagai bakoitzerako bere zehaztasun handiko kategoria aukeratzen da. Enpirikoki erakutsi da existitzen diren ezagutza-baseek eragin positiboa dutela entitateen zehaztasun handiko kategoria ezartzeko asmatze-tasan. Proposatutako sistemak emaitza egonkorrak erakutsi ditu MultiCoNER 2 ataza-partekatuan, adibide urriko kasuetan ere, non baliabide handiko hizkuntzen ezagutza-baseak erabili ditugun.

Sainz et al. (Findings-EMNLP 2023)

Sainz O., Campos J., García-Ferrero I., Etxaniz J., de Lacalle O.L., and Agirre E. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In Bouamor H., Pino J., and Bali K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10776–10787, Singapore, December 2023a. Association for Computational Linguistics.

Laburpena: *Position paper* honetan hizkuntzaren prozesamenduko ebaluazio klasikoa, hau da anotatutako *benchmark*-etan oinarritutakoa, arriskuan dagoela argudiatzen du. Datu-kontaminazio okerrena hizkuntza-eredu handi bat ebaluaziorako erabiliko den datu-multzo baten gainean entrenatzen denean gertatzen da. Arazoaren hedadura ezezaguna da, arazoaren neurketa erreza ez delako. Kontaminazioak ereduen errendimenduen gainestimazio bat eragiten du, batez ere, kontaminazio gabeko ereduekin konparatuz gero. Ondorioak larriak izan daitezke, adibidez, ondorio okerrak dituzten artikuluak argitaratuz konklusio zuzenak dituztenak argitaratzen ez diren bitartean. Artikulu honek datu-kontaminazio maila desberdinak definitzen ditu, eta, komunitate-esfortzu baten alde dei egiten da. Baita datu-kontaminazioa detektatzeko neurri automatiko eta erdi-automatikoen garapena bultzatu, eta kontaminazioaren ondorioz ondorio okerrak atera dituzten artikuluak identifikatzea ere.

Zubillaga et al. (LREC-COLING 2024)

Zubillaga M., **Sainz O.**, Estarrona A., Lopez de Lacalle O., Agirre A. Event Extraction in Basque: Typologically motivated Cross-Lingual Transfer-Learning Analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Lingotto Conference Centre, Torino, Italia, May 2024.

Laburpena: Hizkuntzen arteko ezagutza-transferentzia oso zabaldua dago gertaera-erauzketan baliabide gutxi duten hizkuntzen artean. Zehazki, hizkuntza-eredu eleanitza iturri hizkuntza batean entrenatzea eta beste helburu hizkuntza batean aplikatzeari deritzo. Artikulu honek iturri eta helburu hizkuntzen arteko antzekotasun tipologikoak nola eragiten duen hizkuntzen arteko ezagutza-transferentzian ikertzen du. Euskara hizkuntzan oinarritu gara helburu hizkuntza egokia delako bere hurbileko hizkuntzen desberdintasun tipologikoengatik. Gure ikerketa gertaera erauzketako hiru atazetan erakutsi du iturri eta helburu hizkuntzen artean partekatutako ezaugarri linguistikoek badutela eragina ezagutza-transferentzia kalitatean. 72 hizkuntza pare arteko analisiak erakutsi du token mailako sailkapenean oinarritzen diren atazentzat, entitate eta gertaeren identifikazioa, amankomuneko idazkerak eta ezaugarri morfologikoak dutela eragina kalitate handiko transferentzia lortzeko. Egitura iragarpeneko atazetan aldiz, argumentu erauzketa adibidez, hitzen ordena du garrantzi handiena. Aipatutako aurkikuntzez gain, gure analisiak erakutsi du datu-kopurua handituz gero hizkuntza guztiek ez dutela berdin hobetzen eszenatoki eleanitz batean. Esperimentu guztiak aurrera eramateko EusIE datu-multzoa sortu dugu, Euskaraz dagoen lehendabiziko gertaera-erauzketa datu-multzoa eta *Multilingual Event Extraction* datu-multzoan oinarrituta dagoena. Datu-multzoa eta kodea publikoki atzigarri daude.

Etxaniz et al. (ACL 2024)

Etxaniz J., **Sainz O.**, Perez N., Aldabe I., Rigau G., Agirre E., Ormazabal A., Artetxe M., Soroa A. Latxa: An Open Langauge Model and Evaluation Suite for Basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Laburpena: Latxa aurkezten dugu, Euskarazko hizkuntza-eredu handien familia bat 7 eta 70 mila milioi arteko parametro dituztenak. Latxa sortzeko Llama 2 eredua entrenatzen jarraitu dugu 4,3 milioi dokumentu eta 4,2 mila milioi token dituen Euskarazko corpus berri batean. Kalitate handiko ebaluaziorako datu-multzoen gabeziari aurre egiteko, beste lau aukera-anitzeko ebaluazio datu-multzo aurkezten ditugu: EusProficiency, 5.169 gaitasun azterketa ofizialetatik ateratako galderak; EusReading, 532 irakurmen ulermena ebaluatzeko galderak; EusTrivia, 5 ezagutza arloetako 1.715 *trivia* galderak; eta, EusExams, 16.774 oposizio galderak. Esandako ebaluazio sakonean, Latxa beste eredu irekiak baino hobeto egiten du alde handi batez. Gainera, lehiakorra da GPT-4 Turbo-ren aurka hizkuntzaren gaitasunean eta ulermenean, irakurmenean eta ezagutza intentsiboko atazetan atzetik gelditu arren. Hirurak, Latxa hizkuntza-eredu handien familia, aurre-ikasketarako corpora eta ebaluaziorako datu-multzoak publikoki atzigarri daude lizentzia irekien menpe. Gure ebaluazio ingurunea ahalbidetzen du ikerketa berregingarritasuna baliabide urriko hizkuntzen hizkuntza-eredu handiak sortzeko metodoetan.

1.4 Oinarriak

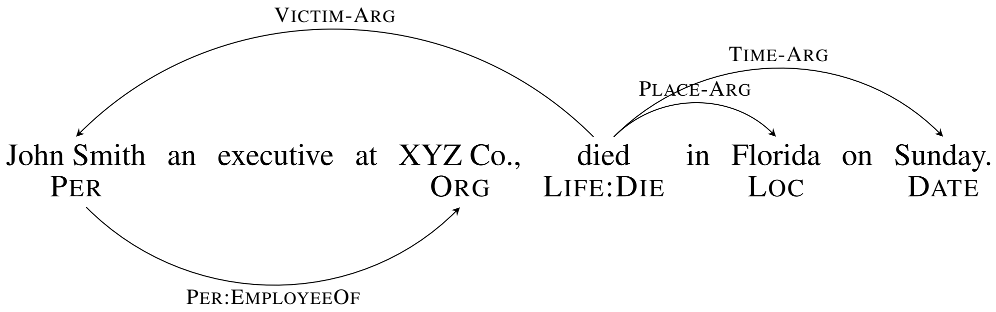
Atal honetan tesia ulertzeko beharrezkoak diren kontzeptu eta aurrekariak azalduko dira. Tesia ikasketa-adibiderik gabeko informazio-erauzketa alorrean kokatzen da. Hori horrela, hasieran informazio-erauzketara sarrera bat egingo da; atazaren motibazioa, helburuak, azpiatazak eta ebaluazio neurriak aurkeztuko dira. Honekin batera, ataza bakoitzaren datu-multzoen bereizitasunak azpimarratuko dira adibideak erabiliz. Ondoren, tesi honentzat garrantzitsua den testuzko inferentzia ataza deskribatuko da, baita atazaren datu-multzo batzuen ezaugarriak erakutsi ere. Azkenik, adibide urriko ikasketaren sarrera bat egingo da. Bertan, hizkuntza-ereduak zer diren azaldu, eta hauek erabiltzen dituzten hurbilpenak azalduko dira.

1.4.1 Informazio-erauzketa

Informazio-erauzketa (IE) zer den azaltzeko, lehendabizi definizio batzuk emango dira. Testu ez-egituratu batetik giza-erabiltzaile batek aurredefinitutako informazio-mota bateko agerpenak datu-egitura batera bihurtzeko helburua duen ataza da informazio-erauzketa (Grishman, 2019). IE sistema baten irteerak egitura jakin bat jarraitu —datu-base bateko sarrera izan daitekeena adibidez— eta beste aplikazio batzuetarako interpretagarria izan behar du. Giza-erabiltzaileak definitutako informazio-mota —edo eskema— erauzi beharreko informazioaren adibideen edota hitzezko adierazpenen bitartez zehaztuta egon daiteke. Giza-erabiltzaileak baliokidetzat identifikatzen dituen bi testu ez-egituratuek datu-egituraren instantzia berera bihurtu beharko lirateke.

Salbuespenak badaude ere, normalean erauzten den informazioa pertsona edo erakunde eta gertaera zehatzetara mugatzen da. Informazio orokorra, baldintzapeko informazioa, ezagutzaren adierazpenak eta sinesmenak kanpoan uzten dira. Murrizketa hauek ataza sinplifikatzearren eta sistemaren irteera hizkuntza-ulermen orokorra baino interpretagarriagoa egitearren aplikatzen dira.

1.1. irudian ikus daiteke gaur egungo informazio-erauzketa atazaren adibide prototipiko bat. Bertan lau ataza —aurrerago azaltzen dira zehatzago— desberdintzen dira: entitate-erauzketa, erlazio-erauzketa, gertaera-erauzketa eta gertaeren argumentu-erauzketa. Entitate eta gertaera-erauzketa atazak sekuentzia-etiketatze moduko atazak dira, aldiz, erlazio eta gertaeren argumentu-erauzketa atazak entitate-entitate edo gertaera-entitate moduko erlazio sailkapen bezala planteatzen dira.


VICTIM-ARG
TIME-ARG
PLACE-ARG
John Smith an executive at XYZ Co., died in Florida on Sunday.
PER
ORG
LIFE:DIE
LOC
DATE
PER:EMPLOYEEOF

1.1 Irudia – Informazio-erauzketan ohikoak diren lau atazekin anotatutako esaldia. Alde batetik PER, ORG, LOC eta DATE entitateak daude markatuta; bestetik, LIFE:DIE gertaera. Entitateen arteko erlazio esaldiaren azpiko geziek adierazten dituzte, eta goiko geziak ordea gertaeren argumentuak.

Hastapenak. Historikoki IE ataza asko aldatu da, batik bat teknologiak aurrera egiten zuen heinean. IEko arloaren hastapenak eta eboluzioa gehien bat Estatu Batuetako gobernuak bultzatutako programen bitartez eraman dira aurrera. MUC (*Message Understanding Conference* ingeleseko sigletatik) programa 1988. urtean hasi zen eta urtero edo bi urtero ebaluaketa orokor bat egiten zen (Sundheim, 1996). Garai hartan, erregeletan oinarritutako metodoak erabiltzen ziren, eta erabiltzen ziren testuak sinpleak baziren ere, erauzi beharreko eskemak konplexuak ziren (40 *slot* zituzten datu-egiturak). Denbora aurrea, ikasketa-automatikoko metodoak garrantzia hartzen joan ziren, eta IEko arloan eragin handia eduki zuten.

Hori horrela, MUC programa bukatu eta ACE (*Automatic Content Extraction* ingeleseko sigletatik) hasi zen 2001. urtean (Walker et al., 2006). Programa honetan gaur egun ezagutzen ditugun ataza-banaketak ezarri ziren, programa honi dagokion datu-multzoa estandarizatuz. ACE, ataza zatiketaz aparte, beste aldaketa batzuk ere ekarri zituen. Alde batetik, ikasketa-datu kopuru handi bat anotatu zen ikasketa-automatikoko sistemak aprobetxatu ahal izateko. Bestetik, entitateen arteko erlazioak edo gertaeren argumentuak esaldi mailara mugatzeko erabakia hartu zen, eta hauek eduki zezaketen *slot* edo argumentu mota kopurua nabarmenki murriztea erabaki zen ere. Segituan nabaritu zen datuen eragina, eta ahalik eta datu gehien erabiltzeko beharra ikusi zen. Hori horrela, eta anotatu gabeko testu kopuru eskala handitzean joan zen heinean metodo erdi-gainbegiratuen garapena bultzatu zen. 2009. urtean TAC-KBP (*Text Analysis Conference*-eko *Knowledge Based Population* ingelesez) ataza hasi zen.

TAC-KBP atazak bi helburu zituen, alde batetik lehen esan bezala hurbilpen erdi-gainbegiratuak bultzatzea anotatutako testu gutxi eta anotatugabeko testu asko emanda; bestetik, dokumentu mailatik haratago joatea, entitateen informazioa jasotzen duen ezagutza-base baten mailako ebaluazioa planteatuz. Ataza honek arrakasta handia izan zuen, batez ere ikasketa-datu urriko informazio-erauzketaren hastapenak ezarri zituelako. Batzuk aipatzeagatik, urruneko-gainbegiraketa (Mintz et al., 2009) eta galdera-erantzun sistemen bidezko ikasketa-adibiderik gabeko IE (Levy et al., 2017). TAC-KBP atazetik TACRED (*TAC Relation Extraction* ingelesezko sigletatik) datu-multzoa sortu zen (Zhang et al., 2017), gaur egungo erlazio-erauzketarako dagoen datu-multzo estandarra.

Ikasketa sakonak eragin handia izan du hizkuntzaren prozesamenduko arloan, eta informazio-erauzketa ez da salbuespen bat izan. Ikasketa sakonak ahalbidetutako eleaniztasunak motibatuta BETTER (*Better Extraction from Text Towards Enhanced Retrieval* ingelesezko sigletatik) programa hasi zen 2019. urtean (Mckinnon and Rubino, 2022). Asmo handiko programa honen helburu nagusia aldibereko informazio-berreskurapen eta erauzketa bultzatzea izan da testuinguru eleanitz batean, hau da, ikasketa hizkuntza batean egin eta beste hizkuntza desberdin batean ebaluatzea. Horretaz aparte, entrenamendurik gabe eskema berri batera egokitzeko gai diren hurbilpenen garapena bultzatu du; tesi hau bertan kokatzen da.

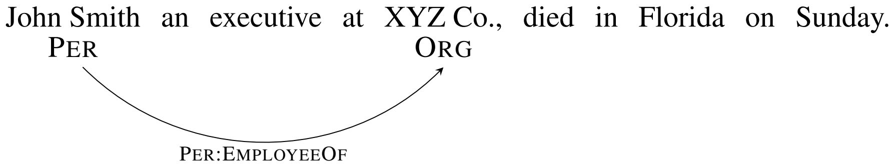
Ataza banaketa MUC-6-rekin hasi eta ACE programarekin estandarizatu ziren. Hurrengo lerroetan ataza bakoitza zertan datzan azalduko da. Oro har, lau dira esaldi mailan aztertzen diren informazio-erauzketa atazak.

- **Entitate-erauzketa** izenak dioen bezala entitateak erauztea du helburu. Entitateak, munduan existitzen diren gizaki, erakunde, leku edo oro har objektu izan daitezke. Salbuespenak egon daitezkeen arren, bi motako entitateak identifika daitezke: entitate izendunak, hau da izen propioa dituzten entitateak ("Fuji mendia" adibidez); eta orokorrak, izenak ez dituztenak baino talde semantiko berekoak direnak ("etxe azpiko parkea" adibidez). Demagun adibide bezala hurrengo esaldia:

John Smith	an	executive	at	XYZ Co.,	died	in	Florida	on	Sunday.
PER				ORG			LOC		DATE

  Adibide honetan "John Smith" pertsona (PER), "XYZ Co." erakundea (ORG), "Florida" lekua (LOC) eta "Sunday" denbora-espresioa (DATE) identifikatzen dira. Kasu batzuetan ohikoa da ere entitateaz aparte balioak ere identifikatzea, adibidez "13€" (MONEY) edo "22 Cº" (TEMPERATURE) modukoak. Ataza hau ataza-katearen oinarrizkoa urratsa izaten da.

- **Erlazio-erauzketa** bi entitate artean dagoen erlazio semantikoa asmatzea, edo erlaziorik ez dagoenean ez dagoela esatea du helburu. Erlazio hauek bi noranzkoak edo noranzko bakarrekoak izan daitezke. Adibidez, TACRED datu-multzoko PER:FAMILY erlazioa bi noranzko erlazio bat da, aldiz datu-multzo bereko PER:PARENTS erlazioa noranzko bakarrekoa da. Hurrengo esaldian:


John Smith an executive at XYZ Co., died in Florida on Sunday.
PER
ORG
PER:EMPLOYEEOF

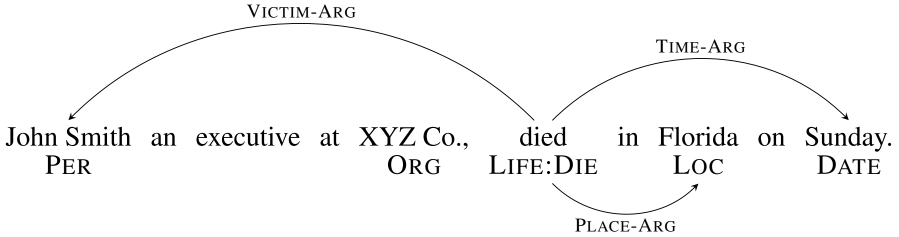
  PER:EMPLOYEEOF erlazioa erakusten da "John Smith"eta "XYZ Co."entitateen artean.

- **Gertaera-erauzketa** testu batean agertzen diren gertaerak identifikatzea eta sailkatzea du helburu. Zehazki, gertaera hobekien adierazten duen testu zatia identifikatzea eta sailkatzea, testu zati horri gertarazlea (*trigger* ingelesez) deitzen zaio. Kasu gehienetan, gertarazlea deiturikoa esaldiko aditza edo aditz nominalizatu bat izango da, baina gerta daiteke ere gertaera famatuen kasuan izen propio baten bitartez adieraztea, "9/11"adibidez. Hurrengo esaldian:

  John Smith an executive at XYZ Co., died in Florida on Sunday.
  LIFE:DIE

  "died" aditzak gertarazten du LIFE:DIE motako gertaera bat.

- **Gertaeren argumentu-erauzketa** testu bat eta bertan agertzen den gertaera bat emanda, gertaeran horren argumentuak edo *slot*-ak identifikatzea eta sailkatzea du helburu. Ataza honetan, gertaerako rol desberdinak identifikatzen dira, normalean nork, zer edo nor, non, noiz eta nola galderei erantzuten dutenak. Gertaera motaren arabera rol hauek izen desberdinak har ditzakete, adibidez hurrengo esaldian:


VICTIM-ARG
TIME-ARG
John Smith an executive at XYZ Co., died in Florida on Sunday.
PER
ORG
LIFE:DIE
LOC
DATE
PLACE-ARG

"John Smith" da nor edo VICTIM-ARG roleko argumentua, "Florida" da non edo PLACE-ARG roleko argumentua eta "Sunday" da noiz edo TIME-ARG roleko argumentua.

Dokumentu mailako edo dokumentu anitzetako informazio-erauzketa. Gaur egungo ikerketa gehienak esaldi mailako informazio-erauzketan oinarritzen dira. Hala ere, informazio-erauzketaren helburua, hasieratik, dokumentu edo dokumentu anitzen gaineko erauzketa burutzea izan da. Esaldi mailara mugitzeko arrazoi nagusia atazaren konplexutasuna txikitzea eta ebaluazioa erraztea izan zen. Hain zuzen ere, dokumentu mailako erauzketaren ebaluaketak, bere horretan, arazoak dituelako. Dokumentu mailako erauzketan, aurredefinitutako eskema bat jarraitzen duten txantiloiak sortzea eta betetzea da helburua. Txantiloi hauek normalean aingura (*anchor* ingelesez) bat izaten dute, hau da, txantiloi hori identifikatzen duen entitate edo gertaera garrantzitsuena —gertaera-erauzketan gertarazlearen antzekoa dena—. Txantiloi horren aingura zein den edota zenbat txantiloi sortu behar diren dokumentu batean erabakitzea ez da erraza, batik bat erreferentzikidetasuna (*coreference* ingelesez) ebatzi behar delako. Ebaluatzerakoan, arazoak agertzen dira ereduak esandako txantiloiak erreferentziakoekin erlazionatzeko garaian. Hori dela eta, alor honetako ikerketa mugatua izan da. Tesi honetan egin diren kontribuzioak ere esaldi mailako informazio-erauzketan egin dira.

Ebaluazioa informazio-erauzketan —esaldi mailan behintzat— nahiko erreza da. Neurri automatikoak erabiltzen dira, zehazki: doitasuna eta estaldura. Hauek kalkulatzeko sistemak zenbat anotazio asmatu dituen (TP), zenbat ez dituen asmatu (FP) eta zenbat ahaztu zaizkion (FN) kontutan hartzen dira. Horrela definitzen dira:

$$\text{doitasuna} = \frac{TP}{TP + FP} \qquad \text{estaldura} = \frac{TP}{TP + FN} \tag{1.1}$$

Informazio-erauzketan adibide bat positiboa dela esaten da anotazio bat esleituta duenean, adibidez, 1.1. irudian ("John Smith", PER) testu-zatia eta ("died", "Florida", LIFE:DIE:PLACE-ARG) tripleta adibide positiboak dira. Aldiz, "an executive" testu-zatia eta ("died", "XYZ Co.") tupla negatiboak dira. Hori horrela, doitasuna definitzen da sistemak esandako adibide positiboen artetik (TP + FP) zenbat diren zuzenak (TP) bezala. Estaldura ordea, benetan positiboak direnen artetik (TP + FN) zenbat diren zuzenak bezala.

Bi neurri edukitzeak sistemen arteko konparaketa zailtzen du, sistema batek neurri batean hobea izateak ez duelako bermatzen bestean ere hobea izatea. Beraz, bi neurri hauek konbinatzen dituzten neurriak erabili izan ohi dira, besteak beste, doitasun-estaldura kurbak edo F1 neurria. Doitasun-estaldura kurba sistemak itzultzen dituen probabilitateen gainean atalase desberdinak ipiniz lortzen diren balioekin osatzen den kurbari deritzo. Kurba honen azpiko azalera ere erabili izan da sistemak konparatzeko garaian. Hala ere, informazio-erauzketan gehien erabiltzen den neurria F1 neurria da, eta doitasun eta estalduraren arteko batezbesteko harmonikoarekin definitzen dena:

$$\mathrm{F1} = \frac{2}{1/\mathrm{doitasuna} + 1/\mathrm{estaldura}} \tag{1.2}$$

Erronkak informazio-erauzketan asko dira. Ataza formulatu zenetik hobekuntza handiak egon diren arren, hizkuntzaren prozesamenduko beste ataza batzuk ez bezala, ebaluazio datu-multzo estandarretan lortzen diren emaitzak perfektu izatetik urrun jarraitzen dute. Gaur egungo metodoek aurre egin behar dieten erronka batzuk hauek dira:

- **Domeinuaren menpekotasuna**. Beste arlo batzuetan bezala, testuaren domeinua eragin handia du informazio-erauzketan. Horren eraginez, joera handia egon da domeinu espezifikoko sistemak garatzeko emaitzak hobeagoak lortzearren. Nola lortu sistema bakar bat domeinu desberdin askotan ondo funtzionaraztea erantzun gabeko galdera bat da oraindik.

- **Anotazio erroreak ebaluazio datuetan.** Asko dira ebaluaziorako erabiltzen diren datu-multzoetan dauden "akatsak". Hain zuzen, anotatzaileen arteko adostasuna oso altua izaten ez delako. Honek eragiten du sistemen emaitzak mugatuta egotea, alegia, anotatzaileen arteko adostasun horretara.

- **Testuinguru falta.** Badaude adibide asko kanpoko ezagutza barik anotatzeko ezinezkoak direnak. Bereziki esanguratsua da esaldi mailako zehaztasun handiko atazetan. Adibidez, demagun hurrengo esaldia: "Peter 1980. urtean jaio zen". Argi dago, bertan "Peter" pertsona —edo izaki bizidun— bati egiten diola erreferentzia, baina zehazki abeslari, jokalari edo bestelako zein zehaztasun handiko entitateri buruz ari den ezinezkoa da jakitea, aurretik Peter nor den jakin gabe behintzat.

- **Datu anotatuen kopuru urria.** Informazio-erauzketan atazak zehaztasun maila handia dute, anotazio gidalerroetan islatzen dena. Ikasketa-automatikoko ereduek zehaztasun hauek ikas ditzaten datuak behar dituzte, eta ez gutxi. Ziur, existitzen diren datu anotatu kopurua bikoiztuz gero emaitzak nabarmenki hobetuko lirateke. Baina, kopurua bikoiztea ez da lan erreza ez merkea. Hori dela eta, metodo ez-/erdi-gainbegiratuak bultzatu dira azken urteotan.

- **Ataza-kate errore propagazioa.** Informazio-erauzketako atazak pausu anitzeko ataza-kateaz daude osatuta, non erroreak hasierako atazetan handiagotzen joaten dira katearen pausu bakoitzean.

Tesi honetan bereziki **datu anotatuen kopuru urriaren** erronkari egin zaio aurre.

1.4.2 Testuzko inferentzia

Testuzko inferentzia —*Textual Entailment* edo *Natural Language Inference* ingelesez— hizkuntzaren ulermena ebaluatzeko erabiltzen den oinarrizko ataza bat da. Dagan et al. (2006)-ek horrela definitzen zuten: izan bitez P eta H, testuzko premisa eta hipotesi bana, P premisatik H hipotesia ondoriozta daitekeen edo ez asmatzean datzan ataza da testuzko inferentzia. Hasierako aldaera honi bi aukerako testuzko inferentzia bezala deitzen zaio. Hain zuzen, P premisa H hipotesia ondorioztatzen duen edo ez erabakien artean aukeratu behar delako soilik. Aurrerago, de Marneffe et al. (2008)-ek aurreko definizioari hirugarren aukera bat gehitu zioten: kontraesana. Hori horrela, gaur egun gehien erabiltzen den atazaren definizioa hurrengoa da: izan bitez P eta H, testuzko premisa eta hipotesi bana, P premisatik H hipotesia ondorioztatzen dela, H hipotesiak P premisa kontraesaten duela edo P premisatik H hipotesia ondorioztatzea ezinezkoa dela aukeren artean erabakitzeko helburua duen ataza da testuzko inferentzia. Erlazio hauek ondorioztapen, kontraesan eta erlazio neutro bezala ezagutzen dira hurrenez-hurren.

1.2. irudian atazaren adibide bat ikus daiteke. Adibide sinple honetan premisaren subjektuarekin jokatu da hipotesi desberdinak egiteko. Lehendabizikoan, ondorioztapen hipotesian, "Ume bat" subjektua "Pertsona bat" subjektuarengatik

Premisa:	***Ume bat****ek pilota jo du.*
Ondorioztapen hipotesi bat:	***Pertsona bat****ek pilota jo du.*
Hipotesi neutro bat:	***Peter****-ek pilota jo du.*
Kontraesan hipotesi bat:	***Heldu bat****ek pilota jo du.*

1.2 Irudia – Testuzko inferentzia atazaren adibidea.

ordezkatu da. Ume bat pertsona bat denez hipotesia premisatik ondoriozta daiteke. Bigarren hipotesian, hipotesi neutroan, subjektua "Peter" izenarengatik ordezkatu da. Kasu honetan, premisako umea Peter den edo ez ezin da jakin, hau da, informazioa falta da hipotesia ondorioztatzeko. Aldi berean, Peter ume bat den edo ez jakin ezin denez ere, premisa ere ezingo litzateke ondorioztatu hipotesitik. Azkenik, heldua izatea ume bat izatearen kontrakoa denez, hipotesiak premisa kontraesaten duela ondoriozta daiteke.

Ataza ebazteko beharrezkoak diren inferentzia gaitasunak asko dira. Aurreko adibidea semantika lexikaleko inferentzian oinarrituta sortu da, baina beste motatako inferentziak existitzen dira ere. Atal honetan inferentzia-mota erabilienak errepasatuko dira.

- **Semantika lexikalean oinarritutako inferentziak** hitzen esanahiean oinarritzen diren inferentziak jasotzen ditu. Adibidez, hitz mailako inferentzia lexikalak: txakur ondorioztatzen du animali baina kontraesaten du katu. Ezeztapen morfologikoko inferentziak: berdin kontraesaten du desberdin. Faktibotasun inferentziak: Nik X ia bukatu dut kontraesaten du Nik X nekez bukatu dut. Simetrikotasun inferentziak: X Y-rekin ezkondu da ondorioztatzen du Y X-rekin ezkondu da, baina X-k Y maite du ez du Y-k X maite du ondorioztatzen. Kopuru inferentziak: gehienak ondorioztatzen du batzuk baina kontraesaten du guztiak.

- **Predikatu-argumentu egituran oinarritutako inferentziak** esaldiaren antolakuntzan oinarritzen diren inferentziak jasotzen ditu. Adibidez aktibo-pasibo aldaerako inferentziak: kanporatu dituzte ondorioztatzen du kanporatuak izan dira. Elipsi inferentziak: etxera korrika etorri nintzen ondorioztatzen du bai etxera etorri nintzen baita korrika etorri nintzen. Besteak beste.

- **Inferentzia logikoak** eragile logikoetan oinarritutako inferentziak jasotzen ditu. Adibidez ezeztapen inferentziak: liburua erosi nuen kontraesaten du liburua ez nuen erosi. Ezeztapen bikoitzeko inferentziak: Ezinezkoa da ez erortzea ondorioztatzen du posiblea da erortzea. Juntagailu inferentziak: gazta eta esnea ekarri ditut ondorioztatzen du bai gazta ekarri dut baita esnea ekarri dut. Disjuntibotasun inferentziak: gazta edo esnea ekarriko dut ez du gazta ekarriko dut ondorioztatzen, baina bai alderantziz. Baldintzapen inferentziak: etortzekotan deituko zaitut ez du deituko zaitut ondorioztatzen.

- **Munduko ezagutzan oinarritutako inferentziak** esplizituki agertzen ez den baina jakintzat ematen den munduari buruzko ezagutzan oinarritutako inferentziak. Adibidez, Donostian badago euskaraz hitz egiten duen jendea ondorioztatzen du Euskal Herrian badago euskaraz hitz egiten duen jendea.

Datu-multzoak. Testuzko inferentzian erabiltzen diren datu-multzoak asko dira batik bat hizkuntza-ereduen ebaluazioan sarri erabiltzen direlako. Atal honetan tesi honetan erabili diren datu-multzoak deskribatuko dira.

- **Stanford Natural Language Inference (SNLI)** datu-multzoa irudien goiburukoen testuetatik sortua izan da (Bowman et al., 2015). Goiburuko hauen testua premisa bezala erabiliz, hipotesiak *crowdsourcing* bitartez sortuak izan dira. Hipotesiak sortzeko hurrengo gidalerroak erabili ziren: idatzi argazkiaren beste goiburuko bat guztiz egiazkoa dena (ondorioztapen hipotesia), idatzi argazkiaren beste goiburuko bat egiazkoa izan daitekeena (hipotesi neutroa); eta, idatzi argazkiaren beste goiburuko bat guztiz faltsua dena (kontraesan hipotesia). Goiburu (premisa) bakoitzeko hainbat hipotesi sortu dira, guztira 570 mila premisa-hipotesi pare sortuz. Oro har, datu-multzo honen premisak eta hipotesiak motzak dira, premisak 47 karaktere bataz-bestean eta hipotesiak berriz 31 karaktere. 1.3. irudian datu-multzo honen adibide bat ikus daiteke.

Premisa:	*A person on a horse jumps over a broken down airplane.*
Ondorioztatzen du:	*A person is outdoors, on a horse.*
Neutroa da:	*A person is training his horse for a competition.*
Kontraesaten dio:	*A person is at a diner, ordering an omelette.*

1.3 Irudia – SNLI datu-multzoko adibidea.

- **Multi-genre Natural Language Inference (MNLI)** datu-multzoa SNLI datu-multzoaren antzera irudien goiburuetatik eratorritako testuzko inferentzia atazaren datu-multzo bat da (Williams et al., 2018). Aurreko datu-multzoaren gidalerro berdinak erabiliz sortu da, eta horren luzapen bat bezala ulertu daiteke. Honen berezitasun nagusia domeinuen dibertsitatea da, aurrekoa ez bezala, datu-multzo hau domeinu desberdineko irudien goiburukoetan oinarritzen da bai entrenamendu baita ebaluazio sendoago bat lortuz. Guztira 433 mila premisa-hipotesi parek osatzen dute datu-multzoa. Kasu honetan, testuak luzeagoak dira: 113 karaktere bataz-bestean premisentzat eta 56 karaktere bataz-bestean hipotesientzat. 1.4. irudian datu-multzo honen adibide bat ikus daiteke.

Premisa:	*This site includes a list of all award winners and a searchable database of Government Executive articles.*
Ondorioztatzen du:	*All of the award winners are listed on the site.*
Neutroa da:	*The Government Executive articles include profiles of notable government employees.*
Kontraesaten dio:	*The Government Executive articles housed on the website are not able to be searched.*

1.4 Irudia – MNLI datu-multzoko adibidea.

- **Fact Extraction and VERification (FEVER)** datu-multzoa datu-egiaztatze atazarako diseinatutako datu-multzo bat da (Thorne et al., 2018). Datu-egiaztatze ataza honetan baieztapen (*claim* ingelesez) eta froga (*evidence* ingelesez) bat emanda, frogak baieztapena eusten duen, ezeztatu egiten duen edota nahiko informazio ez duen asmatzean datza. Ataza hau testuzko inferentzia bezala ebatzi izan da, hau da, frogak baieztapena eusten badu ondorioztatzen dela esaten da, ezeztatu egiten badu kontraesaten dela esaten da eta nahiko informazio ez badu neutrala dela esaten da. Datu-multzoa sortzeko Wikipediako pasarteak erabili dira froga bezala eta baieztapenak eskuz sortu dira. Guztira 248 mila premisa-hipotesi parek osatzen dute datu-multzoa. Datu-multzo honetan, premisak Wikipediako pasarteak direnez luzeagoak dira, 321 karaktere bataz-bestean. Hipotesiak berriz besteak bezain luzeak, 44 karaktere bataz-bestean. 1.5. irudian datu-multzo honen adibide bat ikus daiteke.

- **Adversarial Natural Language Inference (ANLI)** datu-multzoa hizkuntza-ereduentzat zaila izatea du helburu (Nie et al., 2020). Hain zuzen, datu-multzo hau modu aurkaridun batean sortua izan da, hau da, hizkuntza-ereduak egindako erroreetatik eratorria. Sortzeko prozesua iteratiboa izan

Premisa:	*Roman Atwood. He is best known for his vlogs, where he posts updates about his life on a daily basis. His vlogging channel, “RomanAtwoodVlogs”, has a total of 3.3 billion views and 11.9 million subscribers. He also has another YouTube channel called “RomanAtwood”, where he posts pranks.*
Ondorioztatzen du:	*Roman Atwood is a content creator.*

1.5 Irudia – FEVER datu-multzoko adibidea.

da —3 iterazio guztira— eta bakoitzean geroz eta zailagoak diren adibideak sortu dira. Guztira 169 mila premisa-hipotesi parek osatzen dute datu-multzoa. Bataz-bestean premisak 321 karaktere dituzte, hipotesiak aldiz 54. 1.6. irudian datu-multzo honen adibide bat ikus daiteke.

Premisa:	*The Parma trolleybus system (Italian: "Rete filoviaria di Parma") forms part of the public transport network of the city and "comune"of Parma, in the region of Emilia-Romagna, northern Italy. In operation since 1953, the system presently comprises four urban routes.*
Ondorioztatzen du:	*The trolleybus system has over 2 urban routes.*
Neutroa da:	*Since 1953 the Parma Troleybus system has been comprised of four urban routes.*
Kontraesaten dio:	*The Parma trolleybus was widely used in the 1940s.*

1.6 Irudia – ANLI datu-multzoko adibidea.

1.4.3 Adibide urriko ikasketa

Ikasketa-automatikoan, ebatzi beharreko atazari buruzko adibiderik eman gabe —edo gutxi batzuk emanda— ataza egikaritzeko gai diren teknika multzoari adibide urriko ikasketa deritzo. Izan ere, gizakiok gai gara azalpen txiki batekin eta adibide gutxi batzuekin ataza berriak egiten ikasteko; motibazio honetan oinarritzen dira adibide urriko ikasketako teknikak. Lehen aldiz, Chang et al. (2008)-ek proposatu zuten kontzeptu hau, garai hartan daturik gabeko sailkapena edo *dataless classification* bezala deiturikoa.

Adibide urriko ikasketa egikaritzeko, beharrezkoa da aurretik datuen adierazpen semantiko aberats bat edukitzea. Errepresentazio hau aurretik ikasten da, normalean corpus handi baten gainean. Chang et al. (2008)-en kasuan Semantika

Esplizituko Analisia (SEA) erabili zuten adierazpen semantiko bezala, Wikipediaren gainean entrenatuz. SEA —TF-IDF erabiliz lortzen den errepresentazioa— *bag-of-words* moduko adierazpen bat da, non tokenen ordena ez den kontutan hartzen. Hori horrela, hasierako adibide urriko ikasketako teknikak, testu edo irudi sailkapeneko atazetara mugatuta zeuden, hain zuzen, erabiltzen ziren adierazpen semantikoak zituzten mugengatik.

Gaur egun ordea, errepresentazio horiek hizkuntza-eredu handiak erabiliz lortzen dira. *Bag-of-words* moduko adierazpenak ez bezala, hizkuntza-ereduek tokenen —testua errepresentatzeko erabiltzen diren hitz edo azpi-hitz mailako unitate atomikoen— ordena errespetatzen dute, hitzen esanahia ez bakarrik testu osoaren esanahia ere barneratuz.

Atal honetan lehendabizi hizkuntza-ereduei sarrera bat aurkeztuko da. Ondoren, hizkuntza-ereduek adibide urriko ikasketa egikaritzeko tekniken oinarriak azalduko dira, hain zuzen testuzko baldintzapen bidezko ikasketa edo *prompt-learning* zer den azalduko da. Azkenik, pibot-ataza bidezko ikasketara sarrera egingo da, tesi honi dagozkion kontzeptuak azalduz.

Hizkuntza-ereduak (HE) hizkuntzaren adierazpen sakon bat ikasten duten ereduak dira. Hizkuntza-ereduek testu —token sekuentzia— mailan lan egiten dute, token edo n-grametan oinarritutako ereduek ez bezala testuinguruaren esanahia ere barneratuz. Bereziki, azken urteotan Transformer (Vaswani et al., 2017) arkitekturan oinarritutako hizkuntza-ereduak egonkortu dira estandar bat bezala. Hauek hizkuntza-eredu aurre-entrenatuak bezala ere ezagutzen dira, hain zuzen, eredu bat hizkuntza-eredu bihurtzeko aurre-entrenamendu fase batetik igaro behar duelako. Fase honetan, testuzko corpus handi bat erabiliz hizkuntza ulertzeko edota sortzeko —helburu funtzioaren arabera— beharrezkoak diren adierazpenak ikasten dituzte ereduek modu auto-gainbegiratu batean. Helburu funtzio eta arkitektura motaren arabera hizkuntza-eredu desberdinak entrena daitezke:

- **HE kodetzaileak** sarreran emandako testu baten ahalik eta errepresentazio hoberena lortzeko optimizatzen dira. Hori lortu ahal izateko, testuko token bakoitza beste token guztiei arreta jar diezaioke, hau da, bi-noranzkoko arreta-mekanismoa erabiltzen dute. Eredu mota honen eredu esanguratsuenak BERT (Devlin et al., 2019) edo RoBERTa (Liu et al., 2019) dira, besteak beste.

  Eredu hauek entrenatzeko ezkutuko tokenen aurresate (*Masked Token Prediction* edo *Masked Language Modeling* ingelesez) ikasketa helburua erabiltzen da. Labur, testu bat emanda, testuko hainbat token ezkutatzen dira *mask* token berezi bat erabiliz. Gero, hizkuntza-ereduari ezkutatu diren tokenak asmatzeko eskatzen zaio. Modu honetan, ereduari behartzen zaio alboko tokenen informazioaz baliatzea uneko tokenaren errepresentazioa osatzeko.

Eredu mota hau hizkuntzaren ulermenerako atazetan erabilia izan da batez ere, testu-sailkapenean edo sekuentzia-etiketatze moduko atazetan.

- **HE deskodetzaileak** sarreran emandako testu baten jarraipena sortzeko optimizatzen dira. Hori horrela, entrenamendu garaian uneko token bakoitza bere aurrekariei soilik jar diezaioke arreta. Hain zuzen, inferentzia garaian etorkizuneko tokenak momentuan existitzen ez direlako. Erabiltzen duten arreta-mekanismoari kausazko arreta-mekanismoa deitzen zaio. Eredu mota honen eredu esanguratsuenak GPT (Radford and Narasimhan, 2018, Radford et al., 2019, Brown et al., 2020, OpenAI et al., 2024) eta Llama (Touvron et al., 2023a,b) eredu familiak dira.

  Eredu hauek entrenatzeko hurrengo token asmaketan (*Next Token Prediction* ingelesez) ikasketa helburua erabiltzen da. Izenak adierazten duen bezala, token sekuentzia bat emanda, hurrengo tokena zein den asmatzeko eskatzen zaio hizkuntza-ereduari. Honen ondorioz, inferentzia garaian testuzko gonbita bat emanik ereduak koherentea den testua sortzeko ahalmena lortzen du.

  Eredu mota hauek testu sorkuntza atazetan erabili ohi dira, testu laburpenean edo elkarrizketetan besteak beste. Gaur egun, hizkuntza-eredu mota hau erabiliena da, gero ikusiko dugun modura ia edozein ataza testu sorkuntza bezala plantea daitekeelako.

- **Hizkuntza-eredu kodetzaile-deskodetzaileak** sarreran emandako testu batean oinarrituta beste testu bat sortzeko optimizatzen dira. Hauek esentzian beste bi aurrekoen konbinazio bat da, non sarrerako testua zati kodetzaileak prozesatzen duen eta gero irteerako testua deskodetzaileak sortzen duen. Bi zatien konbinazioa arreta-mekanismoan gertatzen da, hain zuzen, deskodetzaileak kasu honetan ez du bakarrik bere tokenei arreta jartzen kodetzailearen tokenei ere. Eredu mota honen eredu esanguratsuenak T5 (Raffel et al., 2020) eta BART (Lewis et al., 2020) ereduak dira.

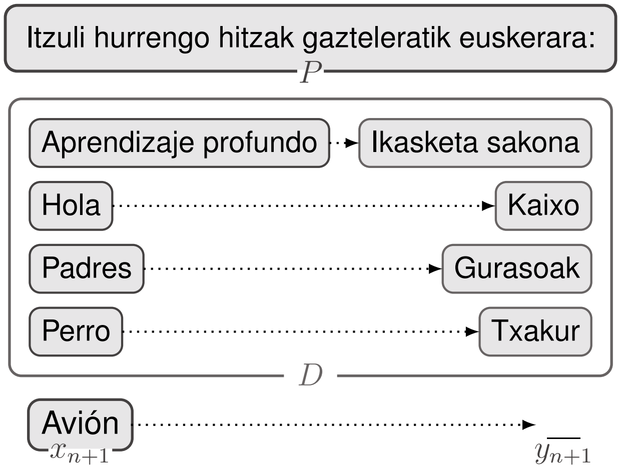
  Eredu hauek entrenatzeko proposamen desberdinak egon dira, famatuena T5 eredu familiak erabiltzen duena da. Eredu kodetzaileen antzera, sarrera testuan token sekuentzia batzuk ordezkatzen dira *sentinel* deituriko token batzuengatik. Gero, deskodetzailea da bertan ezkutatuta zegoen testua asmatzea eskatzen zaiona. Kasu batzuetan, hurrengo token asmaketan ere entrenatzen da deskodetzailea testu naturala sortzeko gaitasuna irabaz dezan.

Eredu mota hauek testutik testurako (*text-to-text* edo *seq-to-seq* ingelesez) bezala ezagutzen diren atazetarako erabiltzen da, itzulpen automatikorako edo testu laburpenerako adibidez. Hala ere, gaur egun eredu deskodetzaileengatik ordezkatuak izan dira gehien bat.

Hizkuntza-eredu moten erabilera asko aldatu da denboran zehar. Hasieran kodetzaileek garrantzi handia hartu zuten testu ulermeneko atazetan, eta, testu sorkuntzako atazak ebatzi behar zirenean kodetzaile-deskodetzaile arkitekturak erabiltzen ziren. Denbora aurrera, gauzak aldatu eta ia ataza guztiak eredu deskodetzaileekin egitera mugitu da arloa. Bi arrazoi izan dira nagusi, alde batetik, edozein ataza testu sorkuntzaren bitartez ebatzi ahal izateko aukerarengatik. Eta bestetik, ereduak handitu ahala eredu deskodetzaileek erakutsi duten propietate emergenteengatik (Wei et al., 2022).

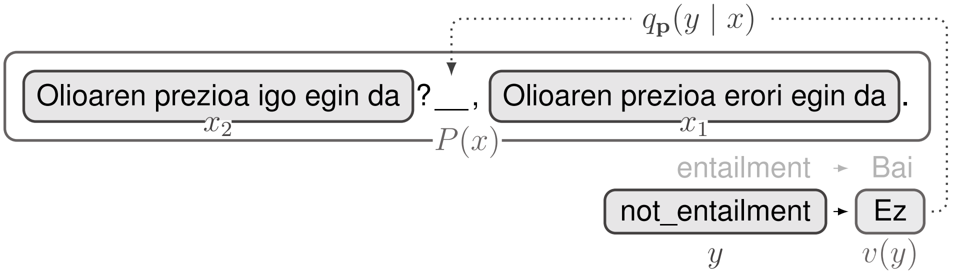
Testuzko baldintzapen bidezko ikasketa. Hizkuntza-ereduen lehendabiziko urteetan ohiko praktika izan zen —eta jarraitzen du izaten— hizkuntza-eredua testu hutsezko corpus handi baten gainean entrenatzea gero ataza espezifiko bateko datuekin eredua fintzeko. Horretarako, hizkuntza-ereduari parametro berri batzuk gehitzen zitzaizkion helburu atazaren formatura egokitzeko. Hau da, hizkuntza-eredua atazara egokitzen zen. Honek hainbat arazo planteatzen zituen adibiderik gabeko ikasketa egikaritzeko, batik bat gehitutako parametro berri horiek findu ezean ereduak ausazko irteerak itzultzen zituelako. Hori horrela, eredua atazara egokitzen zituzten teknikak baino atazak eredura egokitzen zituzten teknikak proposatzen hasi ziren. Teknika guzti horiek testuzko baldintzapen bidezko ikasketaren barnean sailkatzen dira.

- **Testuinguru mailako ikasketa.** GPT-3-ren irteera inflexio puntu bat izan zen, hain zuzen, Brown et al. (2020)-ek erakutsi zutelako ataza testu bezala planteatu eta adibide gutxi batzuk testuinguruan emanda nahikoa zela hizkuntza-eredua ataza hori egiten irakasteko. Hau da, helburu ataza hizkuntza-ereduak dakien atazara —testu sortzera— egokitzea. Teknika honi testuinguru bitarteko ikasketa (*In Context Learning* ingelesez) deitzen zaio. Formalki, $D = \{\{x_0, y_0\}, \{x_1, y_1\}, ...\{x_n, y_n\}\}$ gure entrenamenduko datu-multzo txikia da, non x_i eta y_i i. adibidearen sarrera eta irteera diren. Ereduari eskatzen zaio y_{n+1} irteera asmatzeko, horretarako x_{n+1} sarrera gehi aurreko n adibideak testuinguruan ematen zaizkio. Gainera testuzko baldintzapena P (*prompt* ingelesez) ere gehitzen da ebatzi behar duen ataza azalduz. 1.7. irudian testuinguru bitarteko ikasketaren adibide bat ikus daiteke. Bertan, "*Avión*" gazteleratik euskarara itzultzeko eskatzen zaio, testuinguruan 4 adibide emanda. Teknika hau ahalbidetu zuen GPT-3 bezain handiak diren ereduak atazetara egokitzea eredua findu behar gabe.



1.7 Irudia – Testuinguru mailako ikasketaren adibidea.

- **Txantiloi ustiapen bidezko ikasketa.** Aurreko hurbilpenak hizkuntza-eredu deskodetzaile handiekin funtzionatzen duen arren, ez da horrela eredu kodetzaileen kasuan. Hala ere, helburu ataza hizkuntza-ereduak ikasi duen atazara bihurtzearen ideia eredu hauekin ere aplika daiteke. Hain zuzen hori da Schick and Schütze (2021)-ek proposatzen dutena: txantiloi ustiapen bidezko ikasketa edo PET (*Pattern Exploiting Training* ingelesez). Ideia sinplea da, berriz ere helburu ataza testuz errepresentatzean datza, baina kasu honetan testu irekia sortu beharrean txantiloi bateko hutsuneak betetzeko eskatzen zaio. Formalki, x gure sarrerako testua emanda, $P(x)$ testuzko baldintzapena definitzen da ataza testu bezala adierazteko. Inferentzia egiteko, $\forall y \in Y$ irteeren $v(y)$ hitzezko adierazpenak erabiltzen dira, hizkuntza-ereduari galdetuz zein den adierazpen egokiena $P(x)$ emanda. 1.8. irudian ikus daiteke proposatutako hurbilpenaren adibide bat testuzko inferentzia atazara aplikatuta[2]. Bertan, hizkuntza-eredua baldintzatzen da premisa galdera bezala planteatuz, eta hipotesia erantzunaren parte bihurtuz. Ereduari eskatzen zaio "Bai" —ondorioztatzen du— edo "Ez" —ez du ondorioztatzen— erantzuteko.


$q_{\mathbf{p}}(y \mid x)$
Olioaren prezioa igo egin da
?__,
Olioaren prezioa erori egin da
.
x_2
$P(x)$
x_1
entailment
Bai
not_entailment
Ez
y
$v(y)$

1.8 Irudia – PET sistemaren irudikapena.

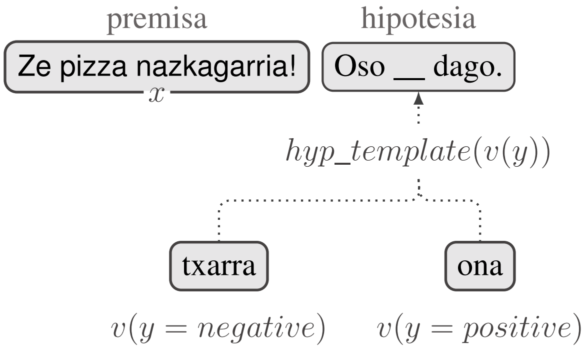
Testuinguru mailako ikasketak ez bezala, hurbilpen honetan adibideak ez dira testuinguruan ematen. Hori horrela, adibide urriko ikasketa egikaritzeko eredua zuzenean fintzen da erabilitako hitzezko adierazpenak erabiliz.

Pibot-atazetan oinarritutako ikasketa. Aurreko hurbilpenetan, testuzko baldintzapen bidezko ikasketan, hizkuntza-ereduak bere horretan erabiltzen dira, beste ezertan entrenatu gabe. Ahaltsuak izan arren, badauzkate hainbat gabezi gainbegiraketarik gabe zailak direnak konpontzeko. Adibidez, hizkuntza-ereduen gaitasunak testuzko inferentzian mugatuak dira adibide gutxi batzuk emanda ere (Brown et al., 2020). Hori horrela, existitzen diren hizkuntzaren ulermenerako atazen datuak hizkuntza-ereduak gainbegiratzeko erabiltzea proposatua izan da. Hain zuzen, horri deritzo pibot-atazetan oinarritutako ikasketa. Erabili izan diren atazak, hizkuntzaren ulermenerako atazen artean entrenamendurako datu aberatsak diren atazak izan dira, besteak beste, galdera-erantzun edo testuzko inferentzia atazak (Min et al., 2023). Hurbilpen honen bitartez adibiderik gabeko ikasketa egikaritzeko prozedura nahiko erreza da. Aurreko metodoetan bezala, egikaritu nahi den ataza —helburu ataza— eredua entrenatua izan den atazaren formatura —iturri edo pibot atazaren formatura— bihurtu behar da. 1.9. irudian erakusten da pibot-atazetan oinarritutako ikasketaren funtzionamenduaren adibide bat, zehazki testuzko inferentzia ataza pibot bezala erabiliz. Adibidean oinarrituz, x sarrera testua premisatzat hartuko da. Txantiloi bidezko ustiapen hurbilpenean bezala, $\forall y \in Y$ irteeren $v(y)$ hitzezko adierazpenak erabiltzen dira. Kasu honetan,

[2] Irudia artikulu originaletik dago aterata eta tesira moldatuta.

adierazpen horiek hipotesiak sortzeko erabiltzen den txantiloi batean txertatzen dira. Inferentzia egiteko, testuzko inferentzia eredu bati eskatzen zaio sortutako hipotesi bakoitzaren ondorioztapen probabilitatea itzultzeko premisa kontutan izanda. Horren arabera hipotesi probableena aukeratzen da irteera bezala.

Berriz ere, txantiloi ustiapen bidezko ikasketa bezala, hurbilpen honek ez ditu testuinguruan adibiderik onartzen. Beraz, adibide urriko ikasketa egikaritzeko eredua zuzenean fintzen da sortutako premisa-hipotesi pareak erabiliz.



1.9 Irudia – Pibot-atazetan oinarritutako ikasketaren adibidea. Zehazki, testuzko inferentzia ataza erabiltzen da pibot bezala.

Tesi honetan pibot-atazetan oinarritutako ikasketa informazio-erauzketa atazetara egokitu da, adibiderik gabeko eta urriko informazio-erauzketa gauzatuz.

1.5 Erlazionatutako lanak

Atal honetan tesiaren gaiarekin erlazionatutako lanak aztertuko dira. Kontuan izan behar da adibide urriko informazio-erauzketaren arloa asko aldatu dela, besteak beste tesi honi eta beste aldibereko edo ondorengo lanei esker. Hori horrela, erlazionatutako lanak arloa jasan dituen faseen arabera antolatu dira. Hasieran informazio-erauzketa gainbegiratuko lanak aztertuko dira 1.5.1. atalean, bertan tesia hasi aurretik informazio-erauzketaren arloaren egoera azalduko da. Ondoren, adibiderik gabeko edo urriko informazio-erauzketan kokatzen diren lanak aurkeztuko dira 1.5.2. atalean, tesi honen aldibereko lanak bertan kokatuz. Azkenik, hizkuntza-eredu handiak erabiltzen dituzten informazio-erauzketako lanak azalduko dira 1.5.3. atalean.

1.5.1 Informazio-erauzketa gainbegiratua

Informazio-erauzketaren arloa, hizkuntzaren prozesamenduko arlo historikoetako bat da, gaur egun 25 urte baino gehiago dituena (Grishman, 2019). Datuak erabiltzen zituzten hurbilpenak ikasketa-automatikoarekin batera etorri ziren. Horien aurretik, hasierako hurbilpenek ez zituzten anotaturiko datuak erabiltzen. Hauek, adituek sortutako ezagutza-baseetan eta erregeletan oinarritzen ziren. Adibidez entitate izendunen erauzketaren kasuan, ezagutza-baseetan oinarritutako metodoak ziren erabilienak (Borkowski and Watson, 1967, Sekine and Nobata, 2004, Etzioni et al., 2005), batez ere bio-medikuntza eta beste domeinu espezifikoetan (Hanisch et al., 2005). Bazeuden ere heuristiko eta beste erregeletan oinarritzen zirenak, "Mr."edo "Ms."ondoren datorrena anotatzea adibidez (Maurel et al., 2011). Erlazio-erauzketaren kasuan ordea erregeletan oinarritutakoak ziren erabilienak (Riloff, 1993, Appelt et al., 1993, Agichtein and Gravano, 2000, Jayram et al., 2006). Gertaera-erauzketan antzeko teknikak erabiltzen ziren ere (Hobbs, 1993, Hobbs et al., 1993). Oro har, metodo hauek datuak behar ez zituzten arren, adituen esfortzu handia beharrezkoa zen hauek inplementatzeko.

Oso azkar ikusi zen eskuz sortutako erregelak eta ezagutza-baseak mugak zituztela, batik bat, hauek sortzeko beharrezkoa zen adituen eskuzko lan neketsuagatik eta orokortzeko zituzten zailtasunengatik. Hau da, metodo horiek doitasun handia lortzen zuten metodoak ziren arren, estaldura aldetik makal gelditzen ziren. Ikasketa-automatikoa aurrera pausu handia izan zen zentzu honetan. Anotaturiko datu-multzo handien sorkuntzarekin batera, ikasketa-automatikoan oinarritutako metodoek garrantzi handia irabazi zuten (Grishman, 2019, Keraghel et al., 2024, Detroja et al., 2023, Lai, 2022). Hurbilpen hauek, eskuz diseinatutako ezaugarriak erabiltzen zituzten datuetatik patroi estatistikoak erauzteko, lehen adituek egin behar zuten lana automatizatuz. Ezaugarri hauek atazaren arabera diseinatzen ziren, adibidez izendun entitateen erauzketan: tokena, bere kategoria gramatikala, tokenaren lema, aurrizkiak eta atzizkiak erabiltzen ziren besteak beste (Bikel et al., 1999, Isozaki and Kazawa, 2002, Kazama et al., 2002, McCallum and Li, 2003, Settles, 2004). Ezaugarri sintaktikoak eta testuingurukoak ere erabili ohi izan dira bai erlazio-erauzketan (Rink and Harabagiu, 2010, Kambhatla, 2004, Zhou et al., 2005) baita gertaeren erauzketan ere (Ahn, 2006, Ji and Grishman, 2008, Patwardhan and Riloff, 2009, Liao and Grishman, 2011).

Ikasketa-automatikoan egon den aurrerapen garrantzitsuenetako bat eta honen bilakaera izan dena ikasketa-sakona da. Ikasketa-sakonak proposatzen duena —hitz xaloetan jarriz— lehen eskuz sortzen ziren ezaugarriak orain algoritmoak ere bere kabuz ikastea da. Bereziki hizkuntzaren prozesamenduan, baita ere informazio-erauzketaren arloan, hitz-bektoreek eragin handia izan zuten. Hitz-bektore edo *embedding* ingelesez, corpus handi batetik modu auto-gainbegiratu batez ikasitako hitzen errepresentazio bektorialei deritzo (Mikolov et al., 2013, Pennington et al., 2014, Bojanowski et al., 2017). Hitz-bektore hauek, testu corpus handiak erabiliz entrenatuak izan direnez, bertan jasotzen den ezagutza errepresentatzea eta baliatzea ahalbidetzen dute. Hauen gainean, arkitektura desberdinak erabili izan ohi dira errepresentazio horiek atazetara egokitzeko. Adibidez, entitate izendunen erauzketan sare-konboluzionalak (Collobert et al., 2011, Li et al., 2017) eta errekurrenteak (Lample et al., 2016, Huang et al., 2015) erabili dira. Erlazio-erauzketan, bereziki sare-konboluzionalak izan dute garrantzi handia (Liu et al., 2013, Zeng et al., 2014, 2015, Nguyen and Grishman, 2015b), baina egon dira sare-errekurrenteak erabili dituztenak ere (Zhang and Wang, 2015, Zhou et al., 2016, Zhan and Zhao, 2020). Azkenik, gertaeren-erauzketan beste atazetan bezala bi arkitekturetan oinarritutako proposamenak egon ziren ere (Nguyen and Grishman, 2015a, Chen et al., 2015, Ghaeini et al., 2016, Kodelja et al., 2019). Garai honetan sare-neuronalen geruza eta arkitektura desberdinen proposamen asko egon ziren, hori horrela, lehendabiziko aldibereko ereduak (*end-to-end* ingelesez) proposatu ziren, non lau atazak aldi berean ebazten ziren (Li et al., 2013, Nguyen et al., 2016).

Hitz-bektore aurre-entrenatuek aurrerapen handiak ekarri zituzten bezala, hurrengo pausu naturala izan den eredu aurre-entrenatuek ere eragin handia izan zuten. Kasu honetan, ez bakarrik ezaugarrien geruza, arkitektura osoa modu auto-gainbegiratu batean entrenatzen baita. Eredu hauei, hizkuntza-eredu deritze. Lehendabiziko hizkuntza-ereduak sare errepikakorretan oinarritutan egon ziren arren (Peters et al., 2018), benetako eraldaketa Transformer (Vaswani et al., 2017) arkitekturan oinarritutako ereduen eskutik etorri zen (Devlin et al., 2019, Liu et al., 2019). Informazio-erauzketan, Transformer kodetzaile arkitekturan oinarritutako ereduak bereziki erabiliak izan dira entitate izendunen erauzketan (Li et al., 2020b, Wang et al., 2021b), erlazio-erauzketan (Baldini Soares et al., 2019, Joshi et al., 2020, Yamada et al., 2020, Wang et al., 2021a, Zhou and Chen, 2022) baita gertaera-erauzketan ere (Wadden et al., 2019, Lin et al., 2020, Nguyen et al., 2021). Azken honetan, proposatzen ziren hurbilpen gehienak dagoeneko aldibereko ereduak izan ziren.

Denbora aurrera joan ahala, eskuz egin beharreko lana murrizten joan da. Lehendabizi, ezagutza-base eta erregelen beharra desagertu zen modu automatiko batean ikasten zituzten algoritmoak erabiltzen hasi zirelako. Ondoren, algoritmo hauek behar zituzten adituek diseinatutako ezaugarriak algoritmoak berak ikasten zituen ezaugarriengatik ordezkatu ziren. Geroz eta gehiago, eskuzko lana datuen sorkuntzara —anotaziora— bideratu da. Lan hau ordea, neketsua eta aditu batek egin beharrekoa izanda, prozesu osoaren zati garestiena bihurtu da. Hurrengo atalean, arazo honi aurre egiten dioten hurbilpenak azalduko dira.

1.5.2 Adibide gabeko eta urriko informazio-erauzketa

Informazio-erauzketa datu anotatu asko behar dituzten atazak osatzen duten arlo bat da. Hori horrela, beti egon da datu hauek merkeago —edo automatikoki— nola lortu ikertzen zuten ikerkerta-lerroak. Baliteke ezagunena urruneko gainbegiraketa izatea da (Mintz et al., 2009). Urruneko gainbegiraketan ezagutza-baseak eta corpus handiak erabiltzen dira automatikoki anotatutako datuak sortzeko. Atazaren arabera aldatu daitekeen arren, atzean dagoen ideia berdina da beti. Entitate izendunen erauzketaren kasuan ezagutza-base batean agertzen diren entitateak corpusean bilatzen dira, gero ezagutza-basean daukaten entitate-mota esleituz. Erlazio-erauzketan egiatzat hartzen da hurrengo premisa: bi entitate ezagutza-base batean erlazionatuta badaude eta bi entitate horiek testuinguru batean batera agertzen badira, testuingurua bien arteko erlazioa erakutsiko duela onartzen da. Modu honetan, ikasketa-automatikoko ereduak datu zaratatsuen gainean entrenatzen dira bertatik ataza ikasteko. Beste lan batzuek ordea, erregeletan oinarritutako sistemak eta corpus handiak erabiliz *bootstraping* teknikak aplikatzea proposatzen dute (Tang and Surdeanu, 2023b,a).

Hizkuntza-ereduen agerpenarekin batera adibide gabeko eta urriko metodoak indarra hartzen hasi ziren (Brown et al., 2020, Schick and Schütze, 2021). Informazio-erauzketaren arloan lehendabiziko datu urriko metodoek erdizka ikusi-gabeko (*partially unseen* ingelesez) ebaluazio protokoloan oinarritzen ziren (Levy et al., 2017, Huang et al., 2018, Obamuyide and Vlachos, 2018, Chen et al., 2020, Du and Cardie, 2020, Li et al., 2020a). Ebaluazio ingurune honetan entrenamendu, garapen eta ebaluazio datu-multzoak entitate, erlazio eta gertaera moten arabera zatitzen ziren, hau da, ebaluazio multzoan dagoen mota bat ezingo da egon ez entrenamendu ezta garapeneko datu-multzoan ere. Datu urriko ikasketaren kasuan N-*way* K-*shot* protokoloa erabiltzen zen (Gao et al., 2019). Ingurune honetan, ebaluazio garaian N mota berrien K adibide ematen ziren erauzketa gauzatzeko.

Hizkuntza-ereduak ahalmentsuagoak bihurtu ziren heinean, beste hizkuntzaren prozesamenduko arloetan bezala lehendabiziko guztiz ikusi-gabeko (*fully-unseen* ingelesez) ikasketa-adibiderik gabeko metodoak garatu ziren. Lehen aipatutako metodoak ez bezala, kasu honetan atazaren adibide bat ere ikusi ez duten metodoei guztiz ikusi-gabeko hurbilpenak deritze. Metodo hauen garapenerako, hizkuntza-ereduen hizkuntza ulermenerako gaitasunak rol garrantzitsu bat izan zuen. Hain zuzen, lehendabiziko metodoek gaitasun honetaz baliatzen zirelako. Tesi honetan garatutako lanak hemen kokatzen dira, testuzko inferentzia gaitasunak baliatuz adibiderik gabeko informazio-erauzketa gauzatzen dutelako (Sainz et al., 2021, 2022a,b). Testuzko inferentzia baliatzen zuten beste lan batzuk egon ziren ere, bai entitate izendunen erauzketan (Li et al., 2022), gertaera-erauzketan (Liu et al., 2020) baita erlazio-erauzketan ere (Zhou et al., 2023, Rahimi and Surdeanu, 2023, Vania et al., 2022). Testuzko inferentziaz aparte beste hizkuntzaren ulermena ebaluatzeko erabiltzen diren atazetan oinarritutako hurbilpenak ere egon ziren; adibidez, galdera-erantzun atazetan (Feng et al., 2020, Liu et al., 2020, Wei et al., 2021, Lyu et al., 2021, Sulem et al., 2022), testu-sorkuntzako atazetan (Du et al., 2021, Li et al., 2021, Lu et al., 2022a, Cui et al., 2021, Lee et al., 2022, Chen et al., 2022) eta beste testu-antzekotasun atazetan (Huang et al., 2022a, Liu et al., 2022, Huang et al., 2022b, Das et al., 2022).

Aurreko metodoak pibot atazetan oinarritutako ikasketa hurbilpen multzoan sailka daitezke, hain zuzen, erdibideko ataza baten gainbegiraketa pibot bezala erabiltzen dutelako helburu ataza ebazteko. Metodo hauek —gaur egun ere artearen egoerarekin lehiakorrak diren arren— baztertuta gelditu dira hizkuntza-eredu handiekin konparatuz. Hurrengo atalean hizkuntza-eredu handiekin informazio-erauzketa gauzatzeko hurbilpenak aztertuko dira.

1.5.3 Hizkuntza-eredu handiak informazio-erauzketan

Hizkuntza-eredu handiek eraldaketa aipagarria izan dute hizkuntzaren prozesamenduan (Brown et al., 2020, OpenAI et al., 2024). Hauen aurretik, arloan egiten zen ikerkuntza atazetara edota ebaluazio datu-multzoetara —*benchmark*-etara— zegoen bideratuta. Orain, ordea, erabilpen kasu zehatzik gabe, orokortasunera bideratuta dagoen ikerkuntza jasotzen du atentzio gehiena (Hendrycks et al., 2021). Hori horrela, gaur egun ez da arraroa ataza bat bakarrik baino, hainbat ataza edota edozein ataza —hobeto edo okerrago— egiteko entrenatuak izan diren ereduak erabiltzea (Chung et al., 2022, Wang et al., 2022). Ondorioz, adibide urriko edo gabeko ebaluazioak salbuespena baino ebaluazio mota lehenetsia bihurtu dira. Aipatzekoa da ere hizkuntza-eredu handien ikerkuntza gehienbat erakunde teknologikoen menpe dagoela, non hizkuntza-ereduak ikerkuntza artefakto baino produktu gisa argitaratzen diren (OpenAI et al., 2024, Touvron et al., 2023b, Anil et al., 2023). Horrek eragin du hizkuntza-eredu handiak erabiltzen dituzten informazio-erauzketako aurrekari askok produktu hauek ebaluatzera mugatzea, oso ondo jakin gabe nola entrenatuak izan diren (Sainz et al., 2023).

Informazio-erauzketa hizkuntza-eredu handiekin gauzatzeko egon ziren lehendabiziko hurbilpenak T5 kodetzaile-deskodetzaile ereduan (Raffel et al., 2020) oinarritu ziren. Lu et al. (2022b) eta Wang et al. (2023a)-ek informazio-erauzketa ataza guztiak testutik-testura (*text-to-text* ingelesez) bihurtzea proposatzen zuten ataza eta datu-multzo askoetatik aldi berean ikasteko. Hala ere, erauzi beharreko eskema errepresentatzeko zehaztasun handirik gabeko adierazpenak erabiltzen zituzten. Informazio-erauzketako datu-multzo batetik bestera entitate, erlazio edo gertaera baten definizioa alda daiteke. Hori horrela, tesi honen azkeneko kontribuzioan arazo honi aurre egiten dion hurbilpen bat proposatu zen. Zehazki, erauzi beharreko eskema errepresentatzerakoan anotazio-gidalerroetan dagoen informazioa txertatzea eta hauek jarraitzea bermatzen duen entrenamendu bat egitea proposatu zen (Sainz et al., 2024). Hori egiteko, sarrera eta irteera Python kode bezala planteatu zen (Wang et al., 2023b) eta hizkuntza-eredu handi deskodetzailetara —Llama 2-ra (Touvron et al., 2023b) adibidez— jauzi egin zen. Li et al. (2024b)-ek gidalerroak hobeto jarraitzeko aurre-entrenamendu bat egitea proposatu zuten ondoren.

Eredu pribatuak erabiltzen dituzten hurbilpenen kasuan, bi dira jarraitzen diren tendentziak: elkarrizketa moduko testuzko baldintzak erabiltzea edo kode moduko testuzko baldintzak erabiltzea. Elkarrizketa bidezko informazio-erauzketaren kasuan, Wei et al. (2024)-ek hainbat pausoko erauzketan oinarritzen den hurbilpen bat proposatzen zuten. Ashok and Lipton (2023)-ek izendun entitateen erauzketan gidalerroak erabiltzea proposatzen zuten, adibiderik gabeko inferentzian arrakasta handirik gabe. Eta, Li et al. (2024a)-ek auto-zuzenketa estrategia bat proposatzen zuen emaitzak hobetzeko. Kode moduko testuzko baldintzak erabiltzen dituzten hurbilpenen kasuan, Wang et al. (2023b) eta Li et al. (2023)-ek izan ziren lehenengoak. Biek kode errepresentazio bat erabili arren, hurbilpen desberdinak proposatu zituzten. Guo et al. (2023)-ek testu berreskurapeneko (*text-retrieval* ingelesez) teknikak txertatzea proposatu zuten.

2. KAPITULUA

Conclusions and Future Work

In this thesis, we have proved that it is possible to perform zero-shot Information Extraction with very competent results. In addition, we have shown that the knowledge transfer between different schemas is feasible, further improving the results in zero-shot. Moreover, we have proposed a new workflow that allows novice users to model complex IE schemas with strong zero-shot performance that later can be curated by expert annotators with much less effort, and, we developed a practical demonstration of the proposed workflow. Finally, we have analyzed and addressed the limitations of the Textual Entailment approach, leveraging the progress carried out by decoder-only LLMs, and implemented GoLLIE: an LLM capable of following more complex annotation guidelines to perform IE annotations. More in detail, the **main contributions** of this thesis are as follows:

- We first proposed to use the Textual Entailment task as a pivot for Information Extraction. Based on the proposal by Yin et al. (2019) for text classification, we adapted and improved the approach for the Relation Extraction task, which we further extended to the rest of the Information Extraction pipeline. The intuition behind the approach is easy: instead of adapting the model for a new task, we adapt our task of interest to another task —the source task— that the model already knows how to solve. In our case, the source task is Textual Entailment. We showed that by approaching Relation Extraction this way, we can obtain zero-shot results similar to supervised results from a few years ago. More importantly, if a couple of examples per relation are given to the model it is able to obtain results close to the supervised state-of-the-art, only 4 F1 points behind. Similar results are obtained for the rest of the tasks. We also performed some distillation and further analyses of our approach.

- We studied the schema-independency property of our proposed approach, that is, whether the model can be trained on a schema and later be evaluated on a different schema. This property is particularly interesting for information extraction tasks, where the schema can vary significantly between different datasets. We evaluated the knowledge transfer between two datasets from the Event Argument Extraction task and showed that a model trained on one schema can improve the results of the zero-shot model when evaluated on another schema. Additionally, we showed that the model can be further trained on the target schema with just a few examples, improving the performance significantly.

- We analyzed the impact of the different strategies when it comes to creating verbalizations. We performed the experiments by asking a second expert —who used a different style— to create hypothesis templates for the same event-argument pairs. The obtained results were positive, as both developers obtained very close results. Additionally, we compared the effort of creating verbalizations for the Event Argument Extraction task with the effort of creating annotations for the same task. We found that the effort of creating verbalizations was significantly lower than the effort of creating the annotations in the amount required to obtain the same results.

- As part of the research project supporting this thesis, we have developed a proof of concept of what we called *verbalize-while-defining* workflow. This workflow is an iterative process that proposes to involve automatic annotations based on verbalizations to facilitate the task of annotation, similar to post-editing machine-translated texts. This process aims to replace the traditional *define, annotate, and train*, where experts first need to define extensive guidelines and annotate large amounts of documents to finally train a model. Based on the premise that zero-shot models work out of the box, it allows novice users to model complex IE schemas with strong zero-shot performance that later can be curated by expert annotators with much less effort. We developed a practical demonstration of the proposed workflow.

- We analyzed and addressed the limitations of the Textual Entailment approach. We leveraged the progress carried out by decoder-only LLMs and implemented GoLLIE: an LLM capable of following annotation guidelines to perform IE annotations. Different from the Textual Entailment approach, GoLLIE leverages detailed guidelines —thanks to the longer context window— instead of simple verbalizations, allowing the model to follow more fine-grained instructions and nuances described by annotation guidelines. Additionally, we provided an error analysis and future research directions for the field.

Regarding the publications, this thesis is comprised of 4 papers that are published in top-tier conferences (1 at EMNLP, 2 at NAACL, and 1 at ICLR). In addition, we also published 7 peer-reviewed papers during the development of this thesis (2 at GWC, 1 at ACM Computing Surveys, 1 at SemEval, 1 at Findings-EMNLP, 1 at LREC-COLING and 1 at ACL).

In terms of software and model artifacts developed during this thesis, they were publicly released under open-source licenses. The code for the Zero-shot Information Extractor is available at the Ask2Transformers[1] GitHub repository. The code for the demonstration of the *verbalize-while-defining* workflow is available at the ZS4IE[2] GitHub repository. The code for GoLLIE is available at the GoLLIE[3] GitHub repository. The models are also available at the Hugging Face[4] model hub.

Moving forward. The field has changed significantly since the beginning of this thesis and continues to change very fast. The initial goal —creating a zero-shot information extraction system that yields decent results— was thought to be hard to achieve at the time of drafting the thesis proposal, however, it was completed a couple of months after the start. The field of zero-shot Information Extraction is now a reality, and the research community is working on improving the results and the methods. With the recent advances in the field, particularly on LLMs, it becomes less clear whether task-specific models are preferred over general case solutions. In both cases, there are many challenges that need to be addressed. Some of the research lines we would like to pursue in the **future** are the following:

- **Estimation of quality and gold standards.** Different from many other tasks in the field, state-of-the-art information extraction systems achieve

[1] https://github.com/osainz59/Ask2Transformers

[2] https://github.com/BBN-E/ZS4IE

[3] https://github.com/hitz-zentroa/GoLLIE

[4] https://huggingface.co/HiTZ

scores far from being perfect. But, so do humans. The annotator agreement —even for highly curated datasets like ACE— is not perfect, and it directly affects the scores that we can expect from the models (Grishman, 2019). Particularly in zero-shot scenarios, where the models lack the data from which to learn those biases and nuances of the specific dataset. Driven by recent advances, the field is moving from static benchmarks towards more human preference-based evaluations. This could be an opportunity to redefine the evaluation and think about possibilities where there is not a single perfect solution or it is not completely known.

- **Effective usage of large context sizes.** Current LLMs support context windows ranging from 4096 to 8192 tokens and beyond. However, most of the current approaches are limited to a single sentence or a couple of sentences. As mentioned in Section 1.4.1, information extraction was originally conceived as a document or multi-document level task. Following the trend of the field, we should move towards more document-level approaches. This would allow the models to leverage the full context window and perform more complex reasoning. In addition, larger context sizes can also be used to provide the models with more detailed guidelines, which should help to improve the quality of the annotations.

Bibliografia

Agichtein E. and Gravano L. Snowball: extracting relations from large plain-text collections. *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, page 85–94, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 158113231X. URL https://doi.org/10.1145/336597.336644.

Ahn D. The stages of event extraction. In Boguraev B., Muñoz R., and Pustejovsky J., editors, *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, 1–8, Sydney, Australia, July 2006. Association for Computational Linguistics. URL https://aclanthology.org/W06-0901.

Anil R., Dai A.M., Firat O., Johnson M., Lepikhin D., Passos A., Shakeri S., Taropa E., Bailey P., Chen Z., Chu E., Clark J.H., Shafey L.E., Huang Y., Meier-Hellstern K., Mishra G., Moreira E., Omernick M., Robinson K., Ruder S., Tay Y., Xiao K., Xu Y., Zhang Y., Abrego G.H., Ahn J., Austin J., Barham P., Botha J., Bradbury J., Brahma S., Brooks K., Catasta M., Cheng Y., Cherry C., Choquette-Choo C.A., Chowdhery A., Crepy C., Dave S., Dehghani M., Dev S., Devlin J., Díaz M., Du N., Dyer E., Feinberg V., Feng F., Fienber V., Freitag M., Garcia X., Gehrmann S., Gonzalez L., Gur-Ari G., Hand S., Hashemi H., Hou L., Howland J., Hu A., Hui J., Hurwitz J., Isard M., Ittycheriah A., Jagielski M., Jia W., Kenealy K., Krikun M., Kudugunta S., Lan C., Lee K., Lee B., Li E., Li M., Li W., Li Y., Li J., Lim H., Lin H., Liu Z., Liu F., Maggioni M., Mahendru A., Maynez J., Misra V., Moussalem M., Nado Z., Nham J., Ni E., Nystrom A., Parrish A., Pellat M., Polacek M., Polozov A., Pope R., Qiao S., Reif E., Richter B., Riley P., Ros A.C., Roy A., Saeta B., Samuel R., Shelby R., Slone A., Smilkov D., So D.R., Sohn D., Tokumine S., Valter D., Vasudevan V., Vodrahalli K., Wang X., Wang P., Wang Z., Wang T., Wieting J., Wu Y., Xu

K., Xu Y., Xue L., Yin P., Yu J., Zhang Q., Zheng S., Zheng C., Zhou W., Zhou D., Petrov S., and Wu Y. Palm 2 technical report, 2023.

Appelt D.E., Hobbs J.R., Bear J., Israel D.J., and Tyson M. Fastus: A finite-state processor for information extraction from real-world text. *International Joint Conference on Artificial Intelligence*, 1993. URL https://api.semanticscholar.org/CorpusID:11268011.

Ashok D. and Lipton Z.C. Promptner: Prompting for named entity recognition, 2023.

Baldini Soares L., FitzGerald N., Ling J., and Kwiatkowski T. Matching the blanks: Distributional similarity for relation learning. In Korhonen A., Traum D., and Màrquez L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1279.

Bikel D.M., Schwartz R., and Weischedel R.M. An algorithm that learns what's in a name. *Machine Learning*, 34(1):211–231, Feb 1999. ISSN 1573-0565. URL https://doi.org/10.1023/A:1007558221122.

Bojanowski P., Grave E., Joulin A., and Mikolov T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. URL https://aclanthology.org/Q17-1010.

Borkowski C. and Watson T.J. An experimental system for automatic recognition of personal titles and personal names in newspaper texts. *COLING 1967 Volume 1: Conference Internationale Sur Le Traitement Automatique Des Langues*, 1967. URL https://aclanthology.org/C67-1023.

Bowman S.R., Angeli G., Potts C., and Manning C.D. A large annotated corpus for learning natural language inference. In Màrquez L., Callison-Burch C., and Su J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL https://aclanthology.org/D15-1075.

Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., and Amodei D. Language models are few-shot learners. In Larochelle H., Ranzato M., Hadsell R., Balcan M., and Lin H., editors, *Advances in Neural Information Processing Systems*, 33 lib., 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chang M.W., Ratinov L., Roth D., and Srikumar V. Importance of semantic representation: dataless classification. *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, page 830–835. AAAI Press, 2008. ISBN 9781577353683.

Chen X., Zhang N., Xie X., Deng S., Yao Y., Tan C., Huang F., Si L., and Chen H. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2778–2788, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. URL https://doi.org/10.1145/3485447.3511998.

Chen Y., Xu L., Liu K., Zeng D., and Zhao J. Event extraction via dynamic multi-pooling convolutional neural networks. In Zong C. and Strube M., editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 167–176, Beijing, China, July 2015. Association for Computational Linguistics. URL https://aclanthology.org/P15-1017.

Chen Y., Chen T., Ebner S., White A.S., and Van Durme B. Reading the manual: Event extraction as definition comprehension. In Agrawal P., Kozareva Z., Kreutzer J., Lampouras G., Martins A., Ravi S., and Vlachos A., editors, *Proceedings of the Fourth Workshop on Structured Prediction for NLP*, 74–83, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.spnlp-1.9.

Chung H.W., Hou L., Longpre S., Zoph B., Tay Y., Fedus W., Li Y., Wang X., Dehghani M., Brahma S., Webson A., Gu S.S., Dai Z., Suzgun M., Chen X., Chowdhery A., Castro-Ros A., Pellat M., Robinson K., Valter D., Narang S., Mishra G., Yu A., Zhao V., Huang Y., Dai A., Yu H., Petrov S., Chi E.H., Dean J., Devlin J., Roberts A., Zhou D., Le Q.V., and Wei J. Scaling instruction-finetuned language models, 2022.

Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., and Kuksa P. Natural language processing (almost) from scratch, 2011.

Cui L., Wu Y., Liu J., Yang S., and Zhang Y. Template-based named entity recognition using BART. In Zong C., Xia F., Li W., and Navigli R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1835–1845, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.findings-acl.161.

Dagan I., Glickman O., and Magnini B. The pascal recognising textual entailment challenge. In Quiñonero-Candela J., Dagan I., Magnini B., and d'Alché Buc F., editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, 177–190, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-33428-6.

Das S.S.S., Katiyar A., Passonneau R., and Zhang R. CONTaiNER: Few-shot named entity recognition via contrastive learning. In Muresan S., Nakov P., and Villavicencio A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6338–6353, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.439.

de Marneffe M.C., Rafferty A.N., and Manning C.D. Finding contradictions in text. In Moore J.D., Teufel S., Allan J., and Furui S., editors, *Proceedings of ACL-08: HLT*, 1039–1047, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL https://aclanthology.org/P08-1118.

Detroja K., Bhensdadia C., and Bhatt B.S. A survey on relation extraction. *Intelligent Systems with Applications*, 19:200244, 2023. ISSN 2667-3053. URL https://www.sciencedirect.com/science/article/pii/S2667305323000698.

Devlin J., Chang M.W., Lee K., and Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein J., Doran C., and Solorio T., editors, *Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1423.

Du X. and Cardie C. Event extraction by answering (almost) natural questions. In Webber B., Cohn T., He Y., and Liu Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 671–683, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.49.

Du X., Rush A., and Cardie C. GRIT: Generative role-filler transformers for document-level event entity extraction. In Merlo P., Tiedemann J., and Tsarfaty R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 634–644, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-main.52.

Etzioni O., Cafarella M., Downey D., Popescu A.M., Shaked T., Soderland S., Weld D.S., and Yates A. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134, 2005. ISSN 0004-3702. URL https://www.sciencedirect.com/science/article/pii/S0004370205000366.

Feng R., Yuan J., and Zhang C. Probing and fine-tuning reading comprehension models for few-shot event extraction, 2020.

Gao T., Han X., Zhu H., Liu Z., Li P., Sun M., and Zhou J. FewRel 2.0: Towards more challenging few-shot relation classification. In Inui K., Jiang J., Ng V., and Wan X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6250–6255, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-1649.

Ghaeini R., Fern X., Huang L., and Tadepalli P. Event nugget detection with forward-backward recurrent neural networks. In Erk K. and Smith N.A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 369–373, Berlin, Germany, August 2016. Association for Computational Linguistics. URL https://aclanthology.org/P16-2060.

Grishman R. Twenty-five years of information extraction. *Natural Language Engineering*, 25(6):677–692, 2019.

Guo Y., Li Z., Jin X., Liu Y., Zeng Y., Liu W., Li X., Yang P., Bai L., Guo J., and Cheng X. Retrieval-augmented code generation for universal information extraction, 2023.

Hanisch D., Fundel K., Mevissen H.T., Zimmer R., and Fluck J. Prominer: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 2005. URL http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-23765-3.

Hendrycks D., Burns C., Basart S., Zou A., Mazeika M., Song D., and Steinhardt J. Measuring massive multitask language understanding. *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Hobbs J.R. The generic information extraction system. *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*, 1993. URL https://aclanthology.org/M93-1009.

Hobbs J.R., Appelt D., Bear J., Israel D., Kameyalna M., and Tyson M. FASTUS: A system for extracting information from text. *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, 1993. URL https://aclanthology.org/H93-1026.

Huang J.Y., Li B., Xu J., and Chen M. Unified semantic typing with meaningful label inference. In Carpuat M., de Marneffe M.C., and Meza Ruiz I.V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2642–2654, Seattle, United States, July 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.naacl-main.190.

Huang L., Ji H., Cho K., Dagan I., Riedel S., and Voss C. Zero-shot transfer learning for event extraction. In Gurevych I. and Miyao Y., editors, *Proceedings* *of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2160–2170, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL https://aclanthology.org/P18-1201.

Huang Y., He K., Wang Y., Zhang X., Gong T., Mao R., and Li C. COPNER: Contrastive learning with prompt guiding for few-shot named entity recognition. In Calzolari N., Huang C.R., Kim H., Pustejovsky J., Wanner L., Choi K.S., Ryu P.M., Chen H.H., Donatelli L., Ji H., Kurohashi S., Paggio P., Xue N., Kim S., Hahm Y., He Z., Lee T.K., Santus E., Bond F., and Na S.H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, 2515–2527, Gyeongju, Republic of Korea, October 2022b. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.222.

Huang Z., Xu W., and Yu K. Bidirectional lstm-crf models for sequence tagging, 2015.

Isozaki H. and Kazawa H. Efficient support vector classifiers for named entity recognition. *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. URL https://aclanthology.org/C02-1054.

Jayram T.S., Krishnamurthy R., Raghavan S., Vaithyanathan S., and Zhu H. Avatar information extraction system. *IEEE Data Eng. Bull.*, 29:40–48, 2006. URL https://api.semanticscholar.org/CorpusID:18293364.

Ji H. and Grishman R. Refining event extraction through cross-document inference. In Moore J.D., Teufel S., Allan J., and Furui S., editors, *Proceedings of ACL-08: HLT*, 254–262, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL https://aclanthology.org/P08-1030.

Joshi M., Chen D., Liu Y., Weld D.S., Zettlemoyer L., and Levy O. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020. URL https://aclanthology.org/2020.tacl-1.5.

Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 178–181, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/P04-3022.

Kazama J., Makino T., Ohta Y., and Tsujii J. Tuning support vector machines for biomedical named entity recognition. *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*, 1–8, Phildadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. URL https://aclanthology.org/W02-0301.

Keraghel I., Morbieu S., and Nadif M. A survey on recent advances in named entity recognition, 2024.

Kodelja D., Besanccon R., and Ferret O. Exploiting a more global context for event detection through bootstrapping. *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I*, page 763–770, Berlin, Heidelberg, 2019. Springer-Verlag. ISBN 978-3-030-15711-1. URL https://doi.org/10.1007/978-3-030-15712-8_51.

Lai V.D. Event extraction: A survey, 2022.

Lample G., Ballesteros M., Subramanian S., Kawakami K., and Dyer C. Neural architectures for named entity recognition. In Knight K., Nenkova A., and Rambow O., editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 260–270, San Diego, California, June 2016. Association for Computational Linguistics. URL https://aclanthology.org/N16-1030.

Lee D.H., Kadakia A., Tan K., Agarwal M., Feng X., Shibuya T., Mitani R., Sekiya T., Pujara J., and Ren X. Good examples make a faster learner: Simple demonstration-based learning for low-resource NER. In Muresan S., Nakov P., and Villavicencio A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2687–2700, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.192.

Levy O., Seo M., Choi E., and Zettlemoyer L. Zero-shot relation extraction via reading comprehension. In Levy R. and Specia L., editors, *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 333–342, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL https://aclanthology.org/K17-1034.

Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., and Zettlemoyer L. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In Jurafsky D., Chai J., Schluter N., and Tetreault J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.703.

Li B., Yin W., and Chen M. Ultra-fine entity typing with indirect supervision from natural language inference. *Transactions of the Association for Computational Linguistics*, 10:607–622, 2022. URL https://aclanthology.org/2022.tacl-1.35.

Li F., Peng W., Chen Y., Wang Q., Pan L., Lyu Y., and Zhu Y. Event extraction as multi-turn question answering. In Cohn T., He Y., and Liu Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, 829–838, Online, November 2020a. Association for Computational Linguistics. URL https://aclanthology.org/2020.findings-emnlp.73.

Li P., Sun T., Tang Q., Yan H., Wu Y., Huang X., and Qiu X. CodeIE: Large code generation models are better few-shot information extractors. In Rogers A., Boyd-Graber J., and Okazaki N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15339–15353, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.855.

Li Q., Ji H., and Huang L. Joint event extraction via structured prediction with global features. In Schuetze H., Fung P., and Poesio M., editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 73–82, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1008.

Li S., Ji H., and Han J. Document-level event argument extraction by conditional generation. In Toutanova K., Rumshisky A., Zettlemoyer L., Hakkani-Tur D., Beltagy I., Bethard S., Cotterell R., Chakraborty T., and Zhou Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 894–908, Online, June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.naacl-main.69.

Li X., Jie Z., Feng J., Liu C., and Yan S. Learning with rethinking: Recurrently improving convolutional neural networks through feedback. *ArXiv*, abs/1708.04483, 2017. URL https://api.semanticscholar.org/CorpusID:13677160.

Li X., Feng J., Meng Y., Han Q., Wu F., and Li J. A unified MRC framework for named entity recognition. In Jurafsky D., Chai J., Schluter N., and Tetreault J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5849–5859, Online, July 2020b. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.519.

Li Y., Ramprasad R., and Zhang C. A simple but effective approach to improve structured language model output for information extraction, 2024a.

Li Z., Zeng Y., Zuo Y., Ren W., Liu W., Su M., Guo Y., Liu Y., Li X., Hu Z., Bai L., Li W., Liu Y., Yang P., Jin X., Guo J., and Cheng X. Knowcoder: Coding structured knowledge into llms for universal information extraction, 2024b.

Liao S. and Grishman R. Acquiring topic features to improve event extraction: in pre-selected and balanced collections. In Mitkov R. and Angelova G., editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 9–16, Hissar, Bulgaria, September 2011. Association for Computational Linguistics. URL https://aclanthology.org/R11-1002.

Lin Y., Ji H., Huang F., and Wu L. A joint neural model for information extraction with global features. In Jurafsky D., Chai J., Schluter N., and Tetreault J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7999–8009, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.713.

Liu C., Sun W., Chao W., and Che W. Convolution neural network for relation extraction. In Motoda H., Wu Z., Cao L., Zaiane O., Yao M., and Wang W., editors, *Advanced Data Mining and Applications*, 231–242, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-53917-6.

Liu F., Lin H., Han X., Cao B., and Sun L. Pre-training to match for unified low-shot relation extraction. In Muresan S., Nakov P., and Villavicencio A., editors, *Proceedings of the 60th Annual Meeting of the Association* *for Computational Linguistics (Volume 1: Long Papers)*, 5785–5795, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.397.

Liu J., Chen Y., Liu K., Bi W., and Liu X. Event extraction as machine reading comprehension. In Webber B., Cohn T., He Y., and Liu Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1641–1651, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.128.

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., and Stoyanov V. Roberta: A robustly optimized bert pretraining approach, 2019.

Lu K., Hsu I.H., Zhou W., Ma M.D., and Chen M. Summarization as indirect supervision for relation extraction. In Goldberg Y., Kozareva Z., and Zhang Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, 6575–6594, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.490.

Lu Y., Liu Q., Dai D., Xiao X., Lin H., Han X., Sun L., and Wu H. Unified structure generation for universal information extraction. In Muresan S., Nakov P., and Villavicencio A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5755–5772, Dublin, Ireland, May 2022b. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.395.

Lyu Q., Zhang H., Sulem E., and Roth D. Zero-shot event extraction via transfer learning: Challenges and insights. In Zong C., Xia F., Li W., and Navigli R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 322–332, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-short.42.

Maurel D., Friburger N., Antoine J.Y., Eshkol I., and Nouvel D. Cascades de transducteurs autour de la reconnaissance des entités nommées. *Revue TAL*

: traitement automatique des langues, 52(1):69–96, 2011. URL https://hal.science/hal-00682805.

McCallum A. and Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 188–191, 2003. URL https://aclanthology.org/W03-0430.

Mckinnon T. and Rubino C. The IARPA BETTER program abstract task four new semantically annotated corpora from IARPA's BETTER program. In Calzolari N., Béchet F., Blache P., Choukri K., Cieri C., Declerck T., Goggi S., Isahara H., Maegaard B., Mariani J., Mazo H., Odijk J., and Piperidis S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3595–3600, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.384.

Mikolov T., Sutskever I., Chen K., Corrado G.S., and Dean J. Distributed representations of words and phrases and their compositionality. In Burges C., Bottou L., Welling M., Ghahramani Z., and Weinberger K., editors, *Advances in Neural Information Processing Systems*, 26 lib. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Min B., Ross H., Sulem E., Veyseh A.P.B., Nguyen T.H., Sainz O., Agirre E., Heintz I., and Roth D. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2), sep 2023. ISSN 0360-0300. URL https://doi.org/10.1145/3605943.

Mintz M., Bills S., Snow R., and Jurafsky D. Distant supervision for relation extraction without labeled data. In Su K.Y., Su J., Wiebe J., and Li H., editors, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/P09-1113.

Nguyen M.V., Lai V.D., and Nguyen T.H. Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks. In Toutanova K., Rumshisky A., Zettlemoyer L., Hakkani-

Tur D., Beltagy I., Bethard S., Cotterell R., Chakraborty T., and Zhou Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 27–38, Online, June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.naacl-main.3.

Nguyen T.H., Cho K., and Grishman R. Joint event extraction via recurrent neural networks. In Knight K., Nenkova A., and Rambow O., editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 300–309, San Diego, California, June 2016. Association for Computational Linguistics. URL https://aclanthology.org/N16-1034.

Nguyen T.H. and Grishman R. Event detection and domain adaptation with convolutional neural networks. In Zong C. and Strube M., editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 365–371, Beijing, China, July 2015a. Association for Computational Linguistics. URL https://aclanthology.org/P15-2060.

Nguyen T.H. and Grishman R. Relation extraction: Perspective from convolutional neural networks. In Blunsom P., Cohen S., Dhillon P., and Liang P., editors, *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 39–48, Denver, Colorado, June 2015b. Association for Computational Linguistics. URL https://aclanthology.org/W15-1506.

Nie Y., Williams A., Dinan E., Bansal M., Weston J., and Kiela D. Adversarial NLI: A new benchmark for natural language understanding. In Jurafsky D., Chai J., Schluter N., and Tetreault J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4885–4901, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.441.

Obamuyide A. and Vlachos A. Zero-shot relation classification as textual entailment. In Thorne J., Vlachos A., Cocarascu O., Christodoulopoulos C., and Mittal A., editors, *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 72–78, Brussels, Belgium, November 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-5511.

OpenAI, Achiam J., Adler S., Agarwal S., Ahmad L., Akkaya I., Aleman F.L., Almeida D., Altenschmidt J., Altman S., Anadkat S., Avila R., Babuschkin I., Balaji S., Balcom V., Baltescu P., Bao H., Bavarian M., Belgum J., Bello I., Berdine J., Bernadett-Shapiro G., Berner C., Bogdonoff L., Boiko O., Boyd M., Brakman A.L., Brockman G., Brooks T., Brundage M., Button K., Cai T., Campbell R., Cann A., Carey B., Carlson C., Carmichael R., Chan B., Chang C., Chantzis F., Chen D., Chen S., Chen R., Chen J., Chen M., Chess B., Cho C., Chu C., Chung H.W., Cummings D., Currier J., Dai Y., Decareaux C., Degry T., Deutsch N., Deville D., Dhar A., Dohan D., Dowling S., Dunning S., Ecoffet A., Eleti A., Eloundou T., Farhi D., Fedus L., Felix N., Fishman S.P., Forte J., Fulford I., Gao L., Georges E., Gibson C., Goel V., Gogineni T., Goh G., Gontijo-Lopes R., Gordon J., Grafstein M., Gray S., Greene R., Gross J., Gu S.S., Guo Y., Hallacy C., Han J., Harris J., He Y., Heaton M., Heidecke J., Hesse C., Hickey A., Hickey W., Hoeschele P., Houghton B., Hsu K., Hu S., Hu X., Huizinga J., Jain S., Jain S., Jang J., Jiang A., Jiang R., Jin H., Jin D., Jomoto S., Jonn B., Jun H., Kaftan T., Łukasz Kaiser, Kamali A., Kanitscheider I., Keskar N.S., Khan T., Kilpatrick L., Kim J.W., Kim C., Kim Y., Kirchner J.H., Kiros J., Knight M., Kokotajlo D., Łukasz Kondraciuk, Kondrich A., Konstantinidis A., Kosic K., Krueger G., Kuo V., Lampe M., Lan I., Lee T., Leike J., Leung J., Levy D., Li C.M., Lim R., Lin M., Lin S., Litwin M., Lopez T., Lowe R., Lue P., Makanju A., Malfacini K., Manning S., Markov T., Markovski Y., Martin B., Mayer K., Mayne A., McGrew B., McKinney S.M., McLeavey C., McMillan P., McNeil J., Medina D., Mehta A., Menick J., Metz L., Mishchenko A., Mishkin P., Monaco V., Morikawa E., Mossing D., Mu T., Murati M., Murk O., Mély D., Nair A., Nakano R., Nayak R., Neelakantan A., Ngo R., Noh H., Ouyang L., O'Keefe C., Pachocki J., Paino A., Palermo J., Pantuliano A., Parascandolo G., Parish J., Parparita E., Passos A., Pavlov M., Peng A., Perelman A., de Avila Belbute Peres F., Petrov M., de Oliveira Pinto H.P., Michael, Pokorny, Pokrass M., Pong V.H., Powell T., Power A., Power B., Proehl E., Puri R., Radford A., Rae J., Ramesh A., Raymond C., Real F., Rimbach K., Ross C., Rotsted B., Roussez H., Ryder N., Saltarelli M., Sanders T., Santurkar S., Sastry G., Schmidt H., Schnurr D., Schulman J., Selsam D., Sheppard K., Sherbakov T., Shieh J., Shoker S., Shyam P., Sidor S., Sigler E., Simens M., Sitkin J., Slama K., Sohl I., Sokolowsky B., Song Y., Staudacher N., Such F.P., Summers N., Sutskever I., Tang J., Tezak N., Thompson M.B., Tillet P., Tootoonchian A., Tseng E., Tuggle P., Turley N., Tworek J., Uribe J.F.C., Vallone A., Vijayvergiya A., Voss C., Wainwright C., Wang J.J., Wang A., Wang B., Ward J., Wei J., Weinmann C., Welihinda A., Welinder P., Weng J., Weng L., Wiethoff M., Willner D., Winter C., Wolrich S., Wong H., Workman L., Wu S., Wu J., Wu M., Xiao K., Xu T., Yoo S., Yu K., Yuan Q., Zaremba W., Zellers R., Zhang C., Zhang M., Zhao S., Zheng T., Zhuang J., Zhuk W., and Zoph B. Gpt-4 technical report, 2024.

Patwardhan S. and Riloff E. A unified model of phrasal and sentential evidence for information extraction. In Koehn P. and Mihalcea R., editors, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 151–160, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/D09-1016.

Pennington J., Socher R., and Manning C. GloVe: Global vectors for word representation. In Moschitti A., Pang B., and Daelemans W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. URL https://aclanthology.org/D14-1162.

Peters M.E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., and Zettlemoyer L. Deep contextualized word representations. In Walker M., Ji H., and Stent A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://aclanthology.org/N18-1202.

Radford A. and Narasimhan K. Improving language understanding by generative pre-training. 2018. URL https://api.semanticscholar.org/CorpusID:49313245.

Radford A., Wu J., Child R., Luan D., Amodei D., and Sutskever I. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., and Liu P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020. ISSN 1532-4435.

Rahimi M. and Surdeanu M. Improving zero-shot relation classification via automatically-acquired entailment templates. In Can B., Mozes M., Cahyawijaya S., Saphra N., Kassner N., Ravfogel S., Ravichander A., Zhao C., Augenstein I., Rogers A., Cho K., Grefenstette E., and Voita L., editors, *Proceedings* *of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, 187–195, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.repl4nlp-1.16.

Riloff E. Automatically constructing a dictionary for information extraction tasks. page 811 – 816, 1993. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-0027709268&partnerID=40&md5=3c5ca8f4bcc84848912d8dc9250f3a44. Cited by: 267.

Rink B. and Harabagiu S. UTD: Classifying semantic relations by combining lexical and semantic resources. In Erk K. and Strapparava C., editors, *Proceedings of the 5th International Workshop on Semantic Evaluation*, 256–259, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/S10-1057.

Sainz O., Campos J., García-Ferrero I., Etxaniz J., de Lacalle O.L., and Agirre E. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In Bouamor H., Pino J., and Bali K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10776–10787, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-emnlp.722.

Sainz O., García-Ferrero I., Agerri R., Lopez de Lacalle O., Rigau G., and Agirre E. GoLLIE: Annotation guidelines improve zero-shot information-extraction. *The Twelfth International Conference on Learning Representations*, Vienna, Austria, May 2024. URL https://openreview.net/forum?id=Y3wpuxd7u9.

Sainz O., Gonzalez-Dios I., Lopez de Lacalle O., Min B., and Agirre E. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In Carpuat M., de Marneffe M.C., and Meza Ruiz I.V., editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, 2439–2455, Seattle, United States, July 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-naacl.187.

Sainz O., Lopez de Lacalle O., Labaka G., Barrena A., and Agirre E. Label verbalization and entailment for effective zero and few-shot relation extraction. In Moens M.F., Huang X., Specia L., and Yih S.W.t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1199–1212, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.92.

Sainz O., Qiu H., Lopez de Lacalle O., Agirre E., and Min B. ZS4IE: A toolkit for zero-shot information extraction with simple verbalizations. In Hajishirzi H., Ning Q., and Sil A., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, 27–38, Hybrid: Seattle, Washington + Online, July 2022b. Association for Computational Linguistics. URL https://aclanthology.org/2022.naacl-demo.4.

Sainz O. and Rigau G. Ask2Transformers: Zero-shot domain labelling with pretrained language models. In Vossen P. and Fellbaum C., editors, *Proceedings of the 11th Global Wordnet Conference*, 44–52, University of South Africa (UNISA), January 2021. Global Wordnet Association. URL https://aclanthology.org/2021.gwc-1.6.

Schick T. and Schütze H. Exploiting cloze-questions for few-shot text classification and natural language inference. In Merlo P., Tiedemann J., and Tsarfaty R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 255–269, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-main.20.

Sekine S. and Nobata C. Definition, dictionaries and tagger for extended named entity hierarchy. In Lino M.T., Xavier M.F., Ferreira F., Costa R., and Silva R., editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/65.pdf.

Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets. In Collier N., Ruch P., and Nazarenko A., editors, *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, 107–110, Geneva, Switzerland, August 28th and 29th 2004. COLING. URL https://aclanthology.org/W04-1221.

Sulem E., Hay J., and Roth D. Yes, no or IDK: The challenge of unanswerable yes/no questions. In Carpuat M., de Marneffe M.C., and Meza Ruiz I.V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1075–1085, Seattle, United States, July 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.naacl-main.79.

Sundheim B.M. The Message Understanding Conferences. *TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996*, 35–37, Vienna, Virginia, USA, May 1996. Association for Computational Linguistics. URL https://aclanthology.org/X96-1006.

Tang Z. and Surdeanu M. Bootstrapping neural relation and explanation classifiers. In Rogers A., Boyd-Graber J., and Okazaki N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 48–56, Toronto, Canada, July 2023a. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-short.5.

Tang Z. and Surdeanu M. It Takes Two Flints to Make a Fire: Multitask Learning of Neural Relation and Explanation Classifiers. *Computational Linguistics*, 49 (1):117–156, 03 2023b. ISSN 0891-2017. URL https://doi.org/10.1162/coli_a_00463.

Thorne J., Vlachos A., Christodoulopoulos C., and Mittal A. FEVER: a large-scale dataset for fact extraction and VERification. In Walker M., Ji H., and Stent A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://aclanthology.org/N18-1074.

Tjong Kim Sang E.F. and De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142–147, 2003. URL https://aclanthology.org/W03-0419.

Touvron H., Lavril T., Izacard G., Martinet X., Lachaux M.A., Lacroix T., Rozière

B., Goyal N., Hambro E., Azhar F., Rodriguez A., Joulin A., Grave E., and Lample G. Llama: Open and efficient foundation language models, 2023a.

Touvron H., Martin L., Stone K., Albert P., Almahairi A., Babaei Y., Bashlykov N., Batra S., Bhargava P., Bhosale S., Bikel D., Blecher L., Ferrer C.C., Chen M., Cucurull G., Esiobu D., Fernandes J., Fu J., Fu W., Fuller B., Gao C., Goswami V., Goyal N., Hartshorn A., Hosseini S., Hou R., Inan H., Kardas M., Kerkez V., Khabsa M., Kloumann I., Korenev A., Koura P.S., Lachaux M.A., Lavril T., Lee J., Liskovich D., Lu Y., Mao Y., Martinet X., Mihaylov T., Mishra P., Molybog I., Nie Y., Poulton A., Reizenstein J., Rungta R., Saladi K., Schelten A., Silva R., Smith E.M., Subramanian R., Tan X.E., Tang B., Taylor R., Williams A., Kuan J.X., Xu P., Yan Z., Zarov I., Zhang Y., Fan A., Kambadur M., Narang S., Rodriguez A., Stojnic R., Edunov S., and Scialom T. Llama 2: Open foundation and fine-tuned chat models, 2023b.

Vania C., Lee G., and Pierleoni A. Improving distantly supervised document-level relation extraction through natural language inference. In Cherry C., Fan A., Foster G., Haffari G.R., Khadivi S., Peng N.V., Ren X., Shareghi E., and Swayamdipta S., editors, *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, 14–20, Hybrid, July 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.deeplo-1.2.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L.u., and Polosukhin I. Attention is all you need. In Guyon I., Luxburg U.V., Bengio S., Wallach H., Fergus R., Vishwanathan S., and Garnett R., editors, *Advances in Neural Information Processing Systems*, 30 lib. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wadden D., Wennberg U., Luan Y., and Hajishirzi H. Entity, relation, and event extraction with contextualized span representations. In Inui K., Jiang J., Ng V., and Wan X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5784–5789, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-1585.

Walker C., Strassel S., Medero J., and Maeda K. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45, 2006. URL https://catalog.ldc.upenn.edu/LDC2006T06.

Wang R., Tang D., Duan N., Wei Z., Huang X., Ji J., Cao G., Jiang D., and Zhou M. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In Zong C., Xia F., Li W., and Navigli R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1405–1418, Online, August 2021a. Association for Computational Linguistics. URL https://aclanthology.org/2021.findings-acl.121.

Wang X., Zhou W., Zu C., Xia H., Chen T., Zhang Y., Zheng R., Ye J., Zhang Q., Gui T., Kang J., Yang J., Li S., and Du C. Instructuie: Multi-task instruction tuning for unified information extraction, 2023a.

Wang X., Li S., and Ji H. Code4Struct: Code generation for few-shot event structure prediction. In Rogers A., Boyd-Graber J., and Okazaki N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3640–3663, Toronto, Canada, July 2023b. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.202.

Wang X., Jiang Y., Bach N., Wang T., Huang Z., Huang F., and Tu K. Improving named entity recognition by external context retrieving and cooperative learning. In Zong C., Xia F., Li W., and Navigli R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1800–1812, Online, August 2021b. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-long.142.

Wang Y., Mishra S., Alipoormolabashi P., Kordi Y., Mirzaei A., Naik A., Ashok A., Dhanasekaran A.S., Arunkumar A., Stap D., Pathak E., Karamanolakis G., Lai H., Purohit I., Mondal I., Anderson J., Kuznia K., Doshi K., Pal K.K., Patel M., Moradshahi M., Parmar M., Purohit M., Varshney N., Kaza P.R., Verma P., Puri R.S., Karia R., Doshi S., Sampat S.K., Mishra S., Reddy A S., Patro S., Dixit T., and Shen X. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Goldberg Y., Kozareva Z., and Zhang Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in*

Natural Language Processing, 5085–5109, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.340.

Wei J., Tay Y., Bommasani R., Raffel C., Zoph B., Borgeaud S., Yogatama D., Bosma M., Zhou D., Metzler D., Chi E.H., Hashimoto T., Vinyals O., Liang P., Dean J., and Fedus W. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Wei K., Sun X., Zhang Z., Zhang J., Zhi G., and Jin L. Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction. In Zong C., Xia F., Li W., and Navigli R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4672–4682, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-long.360.

Wei X., Cui X., Cheng N., Wang X., Zhang X., Huang S., Xie P., Xu J., Chen Y., Zhang M., Jiang Y., and Han W. Chatie: Zero-shot information extraction via chatting with chatgpt, 2024.

Williams A., Nangia N., and Bowman S. A broad-coverage challenge corpus for sentence understanding through inference. In Walker M., Ji H., and Stent A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://aclanthology.org/N18-1101.

Yamada I., Asai A., Shindo H., Takeda H., and Matsumoto Y. LUKE: Deep contextualized entity representations with entity-aware self-attention. In Webber B., Cohn T., He Y., and Liu Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6442–6454, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.523.

Yin W., Hay J., and Roth D. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In Inui K., Jiang J., Ng V., and

Wan X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3914–3923, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-1404.

Zeng D., Liu K., Chen Y., and Zhao J. Distant supervision for relation extraction via piecewise convolutional neural networks. In Màrquez L., Callison-Burch C., and Su J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1753–1762, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL https://aclanthology.org/D15-1203.

Zeng D., Liu K., Lai S., Zhou G., and Zhao J. Relation classification via convolutional deep neural network. In Tsujii J. and Hajic J., editors, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2335–2344, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL https://aclanthology.org/C14-1220.

Zhan J. and Zhao H. Span model for open information extraction on accurate corpus. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 9523–9530, Apr. 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6497.

Zhang D. and Wang D. Relation classification via recurrent neural network, 2015.

Zhang Y., Zhong V., Chen D., Angeli G., and Manning C.D. Position-aware attention and supervised data improve slot filling. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, 35–45, 2017. URL https://nlp.stanford.edu/pubs/zhang2017tacred.pdf.

Zhou G., Su J., Zhang J., and Zhang M. Exploring various knowledge in relation extraction. In Knight K., Ng H.T., and Oflazer K., editors, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 427–434, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/P05-1053.

Zhou K., Qiao Q., Li Y., and Li Q. Improving distantly supervised relation extraction by natural language inference. *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. URL https://doi.org/10.1609/aaai.v37i11.26644.

Zhou P., Shi W., Tian J., Qi Z., Li B., Hao H., and Xu B. Attention-based bidirectional long short-term memory networks for relation classification. In Erk K. and Smith N.A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 207–212, Berlin, Germany, August 2016. Association for Computational Linguistics. URL https://aclanthology.org/P16-2034.

Zhou W. and Chen M. An improved baseline for sentence-level relation extraction. In He Y., Ji H., Li S., Liu Y., and Chang C.H., editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 161–168, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.aacl-short.21.

Glosategia

adibide urriko ikasketa *few-shot learning*

adibiderik gabeko ikasketa *zero-shot learning*

aingura *anchor*

aldibereko *end-to-end*

ataza-kate *pipeline*

aurkari *adversarial*

erreferentzia-kidetasuna *coreference*

eskatu ahala *on-demand*

ezagutza-destilazioa *knowledge-distillation*

fintze *fine-tuning*

gertaerazi *(to) trigger*

gertaerazle *trigger*

hitzezko adierazpen *verbalization*

iturri irekiko *open-source*

kausazko arreta-mekanismoa *causal attention-mechanism*

lan-fluxu *workflow*

pibote bidezko ikasketa	*pivot-based learning*
testuinguru bitarteko ikasketa	*in-context learning*
testutik testurako	*text-to-text*
testuzko baldintza	*prompt*
testuzko inferentzia	*textual entailment*

A. APPENDIX

Original papers

In this appendix, we present the original papers presented in the manuscript of this thesis in the recommended reading order.

Label Verbalization and Entailment for Effective Zero- and Few-Shot Relation Extraction

Oscar Sainz **Oier Lopez de Lacalle** **Gorka Labaka**
Ander Barrena **Eneko Agirre**

HiTZ Basque Center for Language Technologies - Ixa NLP Group
University of the Basque Country (UPV/EHU)
{oscar.sainz, oier.lopezdelacalle, gorka.labaka, ander.barrena, e.agirre}@ehu.eus

Abstract

Relation extraction systems require large amounts of labeled examples which are costly to annotate. In this work we reformulate relation extraction as an entailment task, with simple, hand-made, verbalizations of relations produced in less than 15 minutes per relation. The system relies on a pretrained textual entailment engine which is run as-is (no training examples, zero-shot) or further fine-tuned on labeled examples (few-shot or fully trained). In our experiments on TACRED we attain 63% F1 zero-shot, 69% with 16 examples per relation (17% points better than the best supervised system on the same conditions), and only 4 points short of the state-of-the-art (which uses 20 times more training data). We also show that the performance can be improved significantly with larger entailment models, up to 12 points in zero-shot, giving the best results to date on TACRED when fully trained. The analysis shows that our few-shot systems are especially effective when discriminating between relations, and that the performance difference in low data regimes comes mainly from identifying no-relation cases.

1 Introduction

Given a context where two entities appear, the Relation Extraction (RE) task aims to predict the semantic relation (if any) holding between the two entities. Methods that fine-tune large pretrained language models (LM) with large amounts of labelled data have established the state of the art (Yamada et al., 2020). Nevertheless, due to differing languages, domains and the cost of human annotation, there is typically a very small number of labelled examples in real-world applications, and such models perform poorly (Schick and Schütze, 2021).

As an alternative, methods that only need a few examples (few-shot) or no examples (zero-shot) have emerged. For instance, *prompt based learning* proposes hand-made or automatically learned task and label verbalizations (Puri and Catanzaro, 2019; Schick and Schütze, 2021; Schick and Schütze, 2020) as an alternative to standard fine-tuning (Gao et al., 2020; Scao and Rush, 2021). In these methods, the prompts are input to the LM together with the example, and the language modelling objective is used in learning and inference. In a different direction, some authors reformulate the target task (e.g. document classification) as a *pivot task* (typically question answering or textual entailment), which allows the use of readily available question answering (or entailment) training data (Yin et al., 2019; Levy et al., 2017). In all cases, the underlying idea is to cast the target task into a formulation which allows us to exploit the knowledge implicit in pre-trained LM (prompt-based) or general-purpose question answering or entailment engines (pivot tasks).

Prompt-based approaches are very effective when the label verbalization is given by one or two words (e.g. text classification), as they can be easily predicted by language models, but strive in cases where the label requires a more elaborate description, as in RE. We thus **propose to reformulate RE as an entailment problem**, where the verbalizations of the relation label are used to produce a hypothesis to be confirmed by an off-the-shelf entailment engine.

In our work[1] we have manually constructed verbalization templates for a given set of relations. Given that some verbalizations might be ambiguous (between city of birth and country of birth, for instance) we complemented them with entity type constraints. In order to ensure that the manual work involved is limited and practical in real-world applications, we allowed at most 15 minutes of manual labor per relation. The verbalizations are used as-is for zero-shot RE, but we also recast labelled RE examples as entailment pairs and fine-tune the en-

[1] Code and splits available at: `https://github.com/osainz59/Ask2Transformers`

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1199–1212
November 7–11, 2021. ©2021 Association for Computational Linguistics

tailment engine for few-shot RE.

The results on the widely used TACRED (Zhang et al., 2017) RE dataset in zero- and few-shot scenarios are excellent, well over state-of-the-art systems using the same amount of data. In addition our method scales well with large pre-trained LMs and large amounts of training data, reporting the best results on TACRED to date.

2 Related Work

Textual Entailment. It was first presented by Dagan et al. (2006) and further developed by Bowman et al. (2015) who called it Natural Language Inference (NLI). Given a textual premise and hypothesis, the task is to decide whether the premise entails or contradicts (or is neutral to) the hypothesis. The current state-of-the-art uses large pre-trained LM fine-tuned in NLI datasets (Lan et al., 2020; Liu et al., 2019; Conneau et al., 2020; Lewis et al., 2020; He et al., 2021).

Relation Extraction. The best results to date on RE are obtained by fine-tuning large pre-trained language models equipped with a classification head. Joshi et al. (2020) pretrains a masked language model on random contiguous spans to learn span-boundaries and predict the entire masked span. LUKE (Yamada et al., 2020) further pretrains a LM predicting entities from Wikipedia, and using entity information as an additional input embedding layer. K-Adapter (Wang et al., 2020) fixes the parameters of the pretrained LM and use Adapters to infuse factual and linguistic knowledge from Wikipedia and dependency parsing.

TACRED (Zhang et al., 2017) is the largest and most widely used dataset for RE in English. It is derived from the TAC-KBP relation set, with labels obtained via crowdsourcing. Although alternate versions of TACRED have been published recently (Alt et al., 2020; Stoica et al., 2021), the state of the art is mainly tested in the original version.

Zero-Shot and Few-Shot learning. Brown et al. (2020) showed that task descriptions (*prompts*) can be fed into LMs for task-agnostic and few-shot performance. In addition, (Schick and Schütze, 2020; Schick and Schütze, 2021; Tam et al., 2021) extend the method and allow finetuning of LMs on a variety of tasks. Prompt-based prediction treats the downstream task as a (masked) language modeling problem, where the model directly generates a textual response to a given prompt. The manual generation of effective prompts is costly and requires domain expertise. Gao et al. (2020) provide an effective way to generate prompts for text classification tasks that surpasses the performance of hand picked ones. The approach uses few-shot training with a generative T5 model (Raffel et al., 2020) to learn to decode effective prompts. Similarly, Liu et al. (2021) automatically search prompts in a embedding space which can be simultaneously fine-tuned along with the pre-trained language model. Note that previous prompt-based models run their zero-shot models on a semi-supervised setting in which some amount of labeled data is given in training. Prompts can be easily generated for text classification. Other tasks require more elaborate templates (Goswami et al., 2020; Li et al., 2021) and currently no effective prompt-based methods for RE exist.

Besides prompt-based methods, the use of pivot tasks has been widely use for few/zero-shot learning. For instance, relation and event extraction have been cast as a question answering problem (Levy et al., 2017; Du and Cardie, 2020), associating each slot label to at least one natural language question. Closer to our work, NLI has been shown too to be a successful pivoting task for text classification (Yin et al., 2019, 2020; Wang et al., 2021; Sainz and Rigau, 2021). These works verbalize the labels, and apply an entailment engine to check whether the input text entails the label description.

In similar work to ours, the relation between entailment and RE was explored by Obamuyide and Vlachos (2018). In their work they present some preliminary experiments where they cast RE as entailment, but only evaluate performance as binary entailment, not as a RE task. As a consequence they do not have competing positive labels and avoid RE inference and the issue of detecting no-relation.

Partially vs. fullly unseen labels in RE. Existing zero/few-shot RE models usually see some labels during training (*label partially unseen*), which helps generalize to the unseen label (Levy et al., 2017; Obamuyide and Vlachos, 2018; Han et al., 2018; Chen and Li, 2021). These approaches do not fully address the data scarcity problem. In this work we address the more challenging *label fully unseen* scenario.

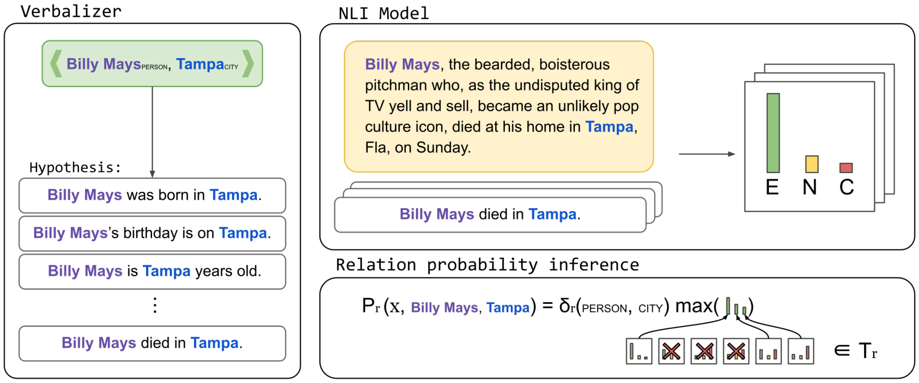

Verbalizer
Billy Mays PERSON, Tampa CITY
Hypothesis:
Billy Mays was born in Tampa.
Billy Mays's birthday is on Tampa.
Billy Mays is Tampa years old.
⋮
Billy Mays died in Tampa.
NLI Model
Billy Mays, the bearded, boisterous pitchman who, as the undisputed king of TV yell and sell, became an unlikely pop culture icon, died at his home in Tampa, Fla, on Sunday.
Billy Mays died in Tampa.
E N C
Relation probability inference
Pr (x, Billy Mays, Tampa) = δr(PERSON, CITY) max( )
∈ Tr

Figure 1: General workflow of our entailment-based RE approach.

3 Entailment for RE

In this section we describe our models for zero- and few-shot RE.

3.1 Zero-shot relation extraction

We reformulate RE as an entailment task: given the input text containing the two entity mentions as the premise and the verbalized description of a relation as hypothesis, the task is to infer if the premise entails the hypothesis according to the NLI model. Figure 1 illustrates the main 3 steps of our system. The first step is focused on relation verbalization to generate the set of hypotheses. In the second we run the NLI model[2] and obtain the entailment probability for each hypothesis. Finally, based on the probabilities and the entity types, we return the relation label that maximizes the probability of the hypothesis, including the NO-RELATION label.

Verbalizing relations as hypothesis. The hypotheses are automatically generated using a set of templates. Each template verbalizes the relation holding between two entity mentions. For instance, the relation PER:DATE_OF_BIRTH can be verbalized with the following template: `{subj}'s birthday is on {obj}`. More formally, given the text x that contains the mention of two entities (x_{e1}, x_{e2}) and template t, the hypothesis h is generated by $\text{VERBALIZE}(t, x_{e1}, x_{e2})$, which substitutes the `subj` and `obj` in the t with the entities x_{e1} and x_{e2}, respectively[3]. Figure 1 shows four verbalizations for the given entity pair.

A relation label can be verbalized by one or more templates. For instance, in addition to the previous template, PER:DATE_OF_BIRTH is also verbalized with `{subj} was born on {obj}`. At the same time, a template can verbalize more than one relation label. For example, `{subj} was born in {obj}` verbalizes PER:COUNTRY_OF_BIRTH and PER:CITY_OF_BIRTH. In order to cope with such ambiguous verbalizations, we added the entity type information to each relation, e.g. `COUNTRY` and `CITY` for each of the relations in the previous example. [4]

We defined a function δ_r for every relation $r \in R$ that checks the entity coherence between the template and the current relation label:

$$\delta_r(e_1, e_2) = \begin{cases} 1 & e_1 \in E_{r1} \land e_2 \in E_{r2} \\ 0 & \text{otherwise} \end{cases}$$

where e_1 and e_2 are the entity types of the first and second arguments, E_{r1} and E_{r2} are the set of allowed types for the first and second entities in relation r. This function is used at inference time, to discard relations that do not match the given types. Appendix C lists all templates and entity type restrictions used in this work.

NLI for inferring relations. In a second step we make use of the NLI model to infer the relation label. Given the text x containing two entities x_{e1}

[2] We describe the NLI models in Section 4.3

[3] Note that the entities are given in a fixed order, that is the relation needs to hold between x_{e1} and x_{e2} in that order; the reverse (x_{e2} and x_{e1}) would be a different example.

[4] Alternatively, one could think on more specific verbalizations, such as `{subj} was born in the city of {obj}` for PER:CITY_OF_BIRTH. In the checks done in the available 15 min. such specific verbalizations had very low recall and were not finally selected.

and x_{e2} the system returns the relation $\hat{r}$ from the set of possible relation labels R with the highest entailment probability as follows:

$$\hat{r} = \arg\max_{r \in R} \mathsf{P}_r(x, x_{e1}, x_{e2}) \tag{1}$$

The probability of each relation P_r is computed as the probability of the hypothesis that yields the maximum entailment probability (Eq. 2), among the set of possible hypothesis. In case the two entities do not match the required entity types, the probability would be zero.

$$\mathsf{P}_r(x, x_{e1}, x_{e2}) = \delta_r(e_1, e_2) \max_{t \in T_r} \mathsf{P}_{NLI}(x, \mathit{hyp})$$
$$\mathit{where}\ \mathit{hyp} = \text{VERBALIZE}(t, x_{e1}, x_{e2}) \tag{2}$$

where P_{NLI} is the entailment probability between the input text and the hypothesis generated by the template verbalizer. Although entailment models return probabilities for entailment, contradiction and neutral, P_{NLI} just makes use of the entailment probability[5]. The right hand-side of Figure 1 shows the application of NLI models and how the probability for each relation, P_r, is computed.

Detection of no-relation. In supervised RE, the NO-RELATION case is taken as an additional label. In our case we examined two approaches.

In **template-based detection** we propose an additional template as if it was yet another relation label, and treated it as another positive relation in Eq. 1. The template for NO-RELATION: `{subj} and {obj} are not related`.

In **threshold-based detection** we apply a threshold $\mathcal{T}$ to P_r in Eq. 2. If none of the relations surpasses the threshold, then our system returns NO-RELATION. On the contrary, the model returns the relation label of highest probability (Eq. 1). When no development data is available, the threshold $\mathcal{T}$ is set to 0.5. Alternatively, we estimate $\mathcal{T}$ using the available development dataset, as described in the experimental part.

3.2 Few-Shot relation extraction

Our system is based on a NLI model which has been pretrained on annotated entailment pairs. When labeled relation examples exist, we can reformulate them as labelled NLI pairs, and use them to fine-tune the NLI model to the task at hand, that is, assigning highest entailment probability to the verbalizations of the correct relation, and assigning low entailment probabilities to the rest of the hypothesis (see Eq. 2).

Given a set of labelled relation examples, we use the following steps to produce labelled entailment pairs for fine-tuning the NLI model. 1) For each **positive** relation example we generate at least one **entailment** instance with the templates that describes the current relation. That is, we generate one or several premise-hypothesis pairs labelled as entailment. 2) For each **positive** relation example we generate one **neutral** premise-hypothesis instance, taken at random from the templates that do not represent the current relation. 3) For each **negative** relation example we generate one **contradiction** example, taken at random from the templates of the rest of relations.

If a template is used for the no-relation case, we do the following: First, for each **no-relation** example we generate one **entailment** example with the no-relation template. Then, for each **positive** relation example we generate one **contradiction** example using the no-relation template.

4 Experimental Setup

In this section we describe the dataset and scenarios we have used for evaluation, how we performed the verbalization process, the different pre-trained NLI models we have used and the state-of-the-art baselines that we compare with.

4.1 Dataset and scenarios

We designed three different low-resource scenarios based on the large-scale TACRED (Zhang et al., 2017) dataset. The full dataset consists of 42 relation labels, including the NO-RELATION label, and each example contains the information about the entity type, among other linguistic information. The scenarios are described in Table 1 and are formed by different splits of the original dataset. We applied a stratified sampling method to keep the original label distribution.

Zero-Shot. The aim of this scenario is the evaluation of the models when no data is available for training. We present two different situations on this scenario: 1) no data is available for development (0% split) and 2) a small development set is available with around 2 examples per relation (1%

[5] The probabilities for relations P_r defined in Eq. 2 are independent from each other, which, in a way, they could be easily extended to multi-label classification task.

		Train (Gold)			Train (Silver)			Development		
		# Pos		# Neg	# Pos		# Neg	# Pos		# Neg
Scenario	Split	mean	total	total	mean	total	total	mean	total	total
Full training	100%	317.4	13013	55112	-	-	-	132.6	5436	17195
Zero-Shot	No Dev	-	-	-	-	-	-	0	0	0
	1% Dev	-	-	-	-	-	-	1.9	54	173
Few-Shot	1%	3.6	130	552	-	-	-	1.9	54	173
	5%	16.3	651	2756	-	-	-	7.0	272	861
	10%	32.6	1302	5513	-	-	-	13.6	544	1721
Data Augment.	0%	0	0	0	246.3	9850	41205	1.9	54	173
	1%	3.6	130	552	246.3	9850	41205	1.9	54	173
	5%	16.3	651	2756	246.3	9850	41205	7.0	272	861
	10%	32.6	1302	5513	246.3	9850	41205	13.6	544	1721

Table 1: Statistics about the dataset scenarios based on TACRED used in the paper, including positive examples per relation, total amount of positive examples and the total amount of negative (no-relation) examples.

split)[6]. In this scenario the models are not allowed to train their own parameters but development data is used to adjust the hyperparameters.

Few-Shot. This scenario presents the challenge of solving the RE task with just a few examples per relation. We present three settings commonly used in few-shot learning (Gao et al., 2020) [7]: around 4 examples per relation (1% of the training data in TACRED), around 16 examples per relation (5%) and around 32 examples per relation (10%). We reduced the development set following the same ratio.

Full Training. In this setting we use all available training and development data.

Data Augmentation. In this scenario we want to test whether a silver dataset produced by running our systems on untagged data can be used to train a supervised relation extraction system (cf. Section 3). In this scenario 75% of the training data in TACRED is set aside as unlabeled data[8], and the rest of the training data is used in different splits (ranging from 1% to 10%). Under this setting we carried out two type of experiments: In the zero-shot experiments (0% in the table) we use our NLI based model to annotate the silver data and then fine-tune the RE model exclusively on the silver data. In the few-shot experiments the NLI model is first fine-tuned with the gold data, then used to annotate the silver data and finally the RE model is fine-tuned over both, silver and gold, annotations.

4.2 Hand-crafted relation templates

We manually created the templates to verbalize relation labels, based on the TAC-KBP guidelines which underlie the TACRED dataset. We limited the time for creating the templates of each relation to less than 15 minutes. Overall, we created 1-8 templates per relation (2 on average) (cf. Appendix C for full list).

The verbalization process consists of generating one or more templates that describe the relation and contain the placeholders `{subj}` and `{obj}`. The developer building the templates was given the task guidelines (brief description of the relation, including one or two examples and the type of the entities) and a NLI model (*roberta-large-mnli* checkpoint). For a given relation, he/she would create a template (or set of templates) and check whether the NLI model is able to output a high entailment probability for the template when applied on the guideline example(s). He/she could run this process for any new template that he/she could come up with. There was no strict threshold involved for selecting the templates, just the intuition of the developer. The spirit was to come up with simple templates quickly, and not to build numerous complex templates or to optimize entailment probabilities.

4.3 Pre-Trained NLI models

For our experiments we tried different NLI models that are publicly available with the Hugging Face Transformers (Wolf et al., 2020) python library.

[6] This setting is comparable to one where the examples in the guidelines are used as development.

[7] The commonly reported value in few-shot scenarios is 16 examples per label. We also added the 3-8 and 32 examples settings in the evaluation.

[8] We use part of the original TACRED dataset to produce silver data in order not to introduce noise coming from different documents and/or pre-processing steps.

		MNLI	No Dev ($\mathcal{T} = 0.5$)			1% Dev		
NLI Model	# Param.	Acc.	Pr.	Rec.	F1	Pr.	Rec.	F1
ALBERT$_{\text{xxLarge}}$	223M	90.8	32.6	**79.5**	46.2	55.2	58.1	56.6 ±1.4
RoBERTa	355M	90.2	32.8	75.5	45.7	58.5	53.1	55.6 ±1.3
BART	406M	89.9	39.0	63.1	48.2	60.7	46.0	52.3 ±1.8
DeBERTa$_{\text{xLarge}}$	900M	91.7	40.3	77.7	53.0	**66.3**	59.7	**62.8** ±1.7
DeBERTa$_{\text{xxLarge}}$	1.5B	91.7	**46.6**	76.1	**57.8**	63.2	**59.8**	61.4 ±1.0

Table 2: Zero-Shot scenario results (Precision, Recall and F1) for our system using several pre-trained NLI models in two settings: no development (default threshold $\mathcal{T}$=0.5), and small development (1% Dev.) for setting $\mathcal{T}$. In the leftmost columns we report the number of parameters and the accuracy in MNLI. For the 1% setting we report the median measures along with the F1 standard deviation in 100 runs.

We tested the following models which implement different architectures, sizes and pre-training objectives and were fine-tuned mainly over the MNLI (Williams et al., 2018) dataset[9]: ALBERT (Lan et al., 2020), RoBERTa (Liu et al., 2019), BART (Lewis et al., 2020) and DeBERTa v2 (He et al., 2021). Table 2 reports the number of parameters of these models. Further details on models can be found in Appendix A.

For each of the scenarios we have tested different models. In zero-shot and full training scenarios we compare all the pre-trained models using the templates described in Section 4.2. For few-shot we used RoBERTa for comparability, as it was used in state-of-the-art systems (cf. Section 4.4), and DeBERTa which is the largest NLI model available on the HUB[10]. Finally, we only tested RoBERTa in data-augmentation experiments.

We ran 3 different runs on each of the experiments using different random seeds. In order to make a fair comparison with state-of-the-art systems (cf section 4.4.), we performed a hyperparameter exploration in the full training scenario, using the resulting configuration also in the zero/few-shot scenarios. We fixed the batch size at 32 for both RoBERTa and DeBERTa, and search the optimum learning-rate among $\{1e^{-6}, 4e^{-6}, 1e^{-5}\}$ on the development set. The best results were obtained using $4e^{-6}$ as learning-rate. For more detailed information refer to the Appendix B.

4.4 State-of-the-art RE models

We compared the NLI approach with the systems reporting the best results to date on TACRED: SpanBERT (Joshi et al., 2020), K-Adapter (Wang et al., 2020) and LUKE (Yamada et al., 2020) (cf. Section 2). In addition, we also report the results obtained by the vanilla RoBERTa baseline proposed by Wang et al. (2020) that serves as a reference for the improvements. We re-trained the different systems on each scenario setting using their publicly available implementations and best performing hyperparameters reported by the authors. All these models have a comparable number of parameters.

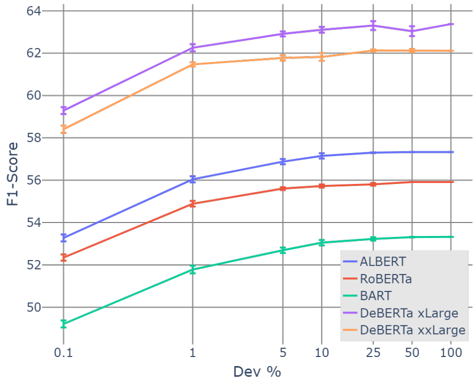


Figure 2: Zero-shot scenario results. Mean F1 and standard error scores when setting $\mathcal{T}$ on increasing number of development examples.

5 Results

5.1 Zero-Shot

Table 2 shows the results for different pre-trained NLI models, as well as the number of parameters and the MNLI *matched* accuracy. These results were obtained by using the threshold for negative relations, as we found that it works substantially better than the no-relation template alternative (cf. Section 3.1). For instance, RoBERTa yields an

[9] ALBERT was trained in some additional NLI datasets.

[10] https://huggingface.co/models

	1%			5%			10%		
Model	Pr.	Rec.	F1	Pr.	Rec.	F1	Prec.	Rec.	F1
SpanBERT	0.0	0.0	0.0 ±0.0	36.3	23.9	28.8 ±13.5	3.2	1.1	1.6 ±20.7
RoBERTa	56.8	4.1	7.7 ±3.6	52.8	34.6	41.8 ±3.3	61.0	50.3	55.1 ±0.8
K-Adapter	73.8	7.6	13.8 ±3.4	56.4	37.6	45.1 ±0.1	62.3	50.9	56.0 ±1.3
LUKE	61.5	9.9	17.0 ±5.9	57.1	47.0	51.6 ±0.4	60.6	60.6	60.6 ±0.4
NLI$_{\text{RoBERTa}}$ (ours)	56.6	55.6	56.1 ±0.0	60.4	68.3	64.1 ±0.2	**65.8**	69.9	67.8 ±0.2
NLI$_{\text{DeBERTa}}$ (ours)	**59.5**	**68.5**	**63.7** ±0.0	**64.1**	**74.8**	**69.0** ±0.2	62.4	**74.4**	**67.9** ±0.5

Table 3: Few-shot scenario results with 1%, 5% and 10% of training data. Precision, Recall and F1 score (standard deviation) of the median of 3 different runs are reported. Top four rows for third-party RE systems run by us.

F1 of 30.1[11] well below the 45.7 when using the default threshold ($\mathcal{T} = 0.5$). Overall we see an excellent zero-shot performance across all the models and settings proving that the approach is robust and model agnostic.

Regarding **pre-trained models**, the best F1 scores are obtained by the two DeBERTa v2 models, which also score the best on the MNLI dataset. Note that all the models achieve similar scores on MNLI, but small differences in MNLI result in large performance gaps when they come to RE, e.g. the 1.5 difference in MNLI between RoBERTa and DeBERTa becomes 7 points in No Dev. and 1% Dev. We think the larger differences in RE are due to the generalization ability of some of the larger models to domain and task differences.

The table includes the results for different values of the $\mathcal{T}$ hyperparameter. In the most challenging setting, with default $\mathcal{T}$, the results are worst, with at most 57.8 F1. However, using as few as 2 examples per relation in average (1% Dev. setting) the results improve significantly.

We performed further experiments using larger amounts of development data to tune $\mathcal{T}$. Figure 2 shows that, for all models, the most significant improvement occurs at the interval [0%, 1%) and that the interval [1%, 100%] is almost flat. The best results with all development data is 63.4%, only 0.6 points better than using 1% of development. These results show clearly that a small number of examples suffice to set an optimal threshold.

5.2 Few-Shot

Table 3 shows the results of competing RE systems and our systems on the few-shot scenario. We report the median and standard deviation across 3 different runs. The competing RE methods suffer a large performance drop, specially for the smallest training setting. For instance, the SpanBERT system (Joshi et al., 2020) has difficulties to converge, even with the 10% of data setting. Both K-Adapter (Wang et al., 2020) and LUKE (Yamada et al., 2020) improve over the RoBERTa system (Wang et al., 2020) in all three settings, but they are well below our NLI$_{\text{RoBERTa}}$ system, with improvements of 48, 22 and 13 points against the baseline in each setting. We also report our method based on DeBERTa$_{\text{xLarge}}$, which is specially effective in the smaller settings.

Model	Pr.	Rec.	F1
SpanBERT	70.8	70.9	70.8
RoBERTa	70.2	72.4	71.3
K-Adapter	70.1	74.0	72.0
LUKE	70.4	75.1	72.7
NLI$_{\text{RoBERTa}}$ (ours)	71.6	70.4	71.0
NLI$_{\text{DeBERTa}}$ (ours)	**72.5**	**75.3**	**73.9**

Table 4: Full training results (TACRED). Top four rows for third-party RE systems as reported by authors.

We would like to note that the zero-shot NLI$_{\text{RoBERTa}}$ system (1% Dev) is comparable in terms of F1 score to a vanilla RoBERTa trained with 10% of the training data. That is, 54 templates (10.5 hours, plus 23 development examples are roughly equivalent to 6800 annotated examples[12] for training (plus 2265 development) .

5.3 Full training

Some zero-shot and few-shot systems are not able to improve results when larger amounts of training data are available. Table 4 reports the results when the whole train and development datasets are used, which is comparable to official results

[11] Results ommitted from Table 2 for brevity.

[12] Unfortunately we could not find the time estimates for annotating examples.

Model	0%	1%	5%	10%
RoBERTa	-	7.7	41.8	55.1
+ Zero-Shot DA	**56.3**	**58.4**	58.8	59.7
+ Few-Shot DA	-	**58.4**	**64.9**	**67.7**

Table 5: Data Augmentation scenario results (F1) for different gold training sizes. Silver annotations by the zero-shot and few-shot NLI$_{\text{RoBERTa}}$ model.

on TACRED. Focusing on our NLI$_{\text{RoBERTa}}$ system, and comparing it to the results in Table 3, we can see that it is able to effectively use the additional training data, improving from 67.9 to 71.0. When compared to a traditional RE system, it performs on a par to RoBERTa, and a little behind K-Adapter and LUKE, probably due to the infused knowledge which our model is not using. These results show that our model keeps improving with additional data and that it is competitive when larger amounts of training is available. The results of NLI$_{\text{DeBERTa}}$ show that our model can benefit from larger and more effective pre-trained NLI systems even in full training scenarios, and in fact achieves the best results to date on the TACRED dataset.

5.4 Data augmentation results

In this section we explore whether our NLI-based system can produce high-quality silver data which can be added to a small amount of gold data when training a traditional supervised RE system, e.g. the RoBERTa baseline (Wang et al., 2020). Table 5 reports the F1 results on the data augmentation scenario for different amounts of gold training data. Overall, we can see that both our zero-shot and few-shot methods[13] provide good quality silver data, as they improve significantly over the baseline in all settings. Although the zero-shot and few-shot methods yield the same result with 1% of training data, the few-shot model is better in the rest of training regimes, showing that it can effectively use the available training data in each case to provide better quality silver data. If we compare the results in this table with those of the respective NLI-based system with the same amount of gold training instances (Tables 2 and 3) we can see that the results are comparable, showing that our NLI-based system and a traditional RE system trained with silver annotations have comparable performance. A practical advantage of a traditional RE system trained with our silver data is that is easier to integrate on available pipelines, as one just needs to download the trained Transformer model. It also makes it easy to check additive improvements in the RE method.

Model	Scenario		P	PvsN
NLI$_{\text{DeBERTa}}$	Zero-Shot	No Dev	85.6	59.5
		1% Dev	85.6	67.7
	Few-Shot	5%	89.7	74.5
	Full train	-	92.2	77.8
LUKE	Few-Shot	5%	69.3	63.4
	Full train	-	90.2	77.3

Table 6: Performance of selected systems and scenarios on two metrics: the binary task of detecting a positive relation vs. no-relation (PvsN column, F1) and detecting the correct relation among positive cases (P, F1).

6 Analysis

Relation extraction can be analysed according to two auxiliary metrics: the binary task of detecting a positive relation vs. no-relation, and the multi-class problem of detecting which relation holds among positive cases (that is, discarding no-relation instances from test data). Table 6 shows the results of a selection of systems and scenarios. The first rows compare the performance of our best system, NLI$_{\text{DeBERTa}}$, across four scenarios, while the last two rows show the results for LUKE in two scenarios. The zero-shot No dev. system is very effective when discriminating the relation among positive examples (P column), only 7 points below the fully trained system, while it lags well behind when discriminating positive vs. negative, 18 points. The use of a small development data for tuning the $\mathcal{T}$ threshold closes the gap in PvsN, as expected, but the difference is still 10 points. All in all, these numbers show that our zero-shot system is very effective discriminating among positive examples, but that it still lags behind when detecting no-relation cases. Overall, the figures show the effectiveness of our methods in low data scenarios on both metrics.

Confusion analysis In supervised models some classes (relations) are better represented in training than others, usually due to data imbalance. Our system instead, represents each relations as a set of templates, which at least on a **zero-shot**

[13] The zero-shot 1% Dev model is used in all data augmentation experiments, while the few-shot method changes to use the available data at each run (1%, 5% and 10%), both with RoBERTa

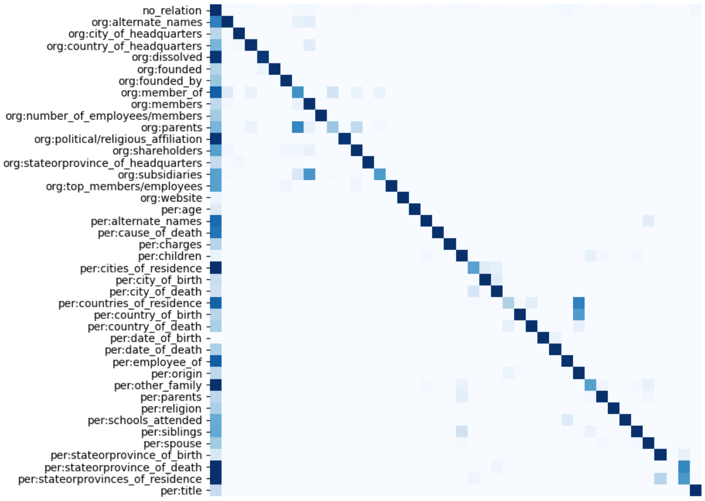


Figure 3: Confusion matrix of our $\text{NLI}_{\text{DeBERTa}}$ zero-shot system on the development dataset. The rows represent the true labels and the columns the predictions. The matrix is rowise normalized (recall in the diagonal).

scenario, should not be affected by data imbalance. The strong diagonal in the confusion matrix (Fig. 3) shows that our the model is able to discriminate properly between most of the relations (after all it achieves 85.6% accuracy, cf. Table 6), with exception of the no-relation column, which was expected. Regarding the confusion between actual relations, most of them are about **overlapping relations**, as expected. For instance, ORG:MEMBER_OF and ORG:PARENTS both involve some organization A being part or member of some other organization B, where ORG:MEMBERS is different from ORG:PARENTS in that correct fillers are distinct entities that are generally capable of autonomously ending their membership with the assigned organization[14]. Something similar occurs between ORG:MEMBERS and ORG:SUBSIDIARIES. Another reason for confusion happens when **two or more relations exist concurrently**, as in PER:ORIGIN, PER:COUNTRY_OF_BIRTH and PER:COUNTRY_OF_RESIDENCE. Finally, the model scores low on PER:OTHER_FAMILY, which is a bucket of many specific relations where only a handful were actually covered by the templates.

7 Conclusions

In this work we reformulate relation extraction as an entailment problem, and explore to what extent simple hand-made verbalizations are effective. The creation of templates is limited to 15 minutes per relation, and yet allows for excellent results in zero- and few-shot scenarios. Our method makes effective use of available labeled examples, and together with larger LMs produces the best results on TACRED to date. Our analysis indicates that the main performance difference against supervised models comes from discriminating no-relation examples, as the performance among positive examples equals that of the best supervised system using the full training data. We also show that our method can be used effectively as a data-augmentation method to provide additional labeled examples. For the future we would like to investigate better methods for detecting no-relation in zero-shot settings.

Acknowledgements

Oscar is funded by a PhD grant from the Basque Government (PRE_2020_1_0246). This work is based upon work partially supported via the IARPA BETTER Program contract No. 2019-19051600006 (ODNI, IARPA), and by the Basque Government (IXA excellence research group IT1343-19).

[14] Description extracted from the guidelines.

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Chih-Yao Chen and Cheng-Te Li. 2021. Zs-bert: Towards zero-shot relation extraction with attribute representation learning. In *Proceedings of 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2021)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners.

Ankur Goswami, Akshata Bhat, Hadar Ohana, and Theodoros Rekatsinas. 2020. Unsupervised relation extraction from language models using constrained cloze completion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1263–1276, Online. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

Raul Puri and Bryan Catanzaro. 2019. Zero-shot text classification with generative language models.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Oscar Sainz and German Rigau. 2021. Ask2Transformers: Zero-shot domain labelling with pretrained language models. In *Proceedings of the 11th Global Wordnet Conference*, pages 44–52, University of South Africa (UNISA). Global Wordnet Association.

Teven Le Scao and Alexander M. Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners. *Computing Research Repository*, arXiv:2009.07118.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shotcomings of the tacred dataset. In *Proceedings of the Thirty-fifth AAAI Conference on Aritificial Intelligence 2021*.

Derek Tam, Rakesh R Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. 2020. Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8229–8239, Online. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45.

A Pre-Trained models

The pre-trained NLI models we have tested from the Transformers library are the next:

- ALBERT: *ynie/albert-xxlarge-v2-snli_mnli _fever_anli_R1_R2_R3-nli*
- RoBERTa: *roberta-large-mnli*
- BART: *facebook/bart-large-mnli*
- DeBERTa v2 xLarge: *microsoft/deberta-v2-xlarge-mnli*
- DeBERTa v2 xxLarge: *microsoft/deberta-v2-xxlarge-mnli*

B Experimental details

We carried out all the experiments on a single Titan V (16GB) except for the fine-tuning of DeBERTa, that has been done on a cluster of 4 Titan V100 (32GB). The average inference time for the zero and few-shot experiments is between 1h and 1.5h. The time needed for fine-tuning the NLI systems was at most 2.5h for RoBERTa and 5h for DeBERTa. All the experiments were done with mixed precision to speed up the overall runtime.

The whole hyperparameter settings used for fine-tuning $\text{NLI}_{\text{RoBERTa}}$ and $\text{NLI}_{\text{DeBERTa}}$ are listed below:

- **Train epochs:** 2
- **Warmup steps:** 1000
- **Learning-rate:** 4e-6
- **Batch-size:** 32
- **FP16 training**
- **Seeds:** {0, 24, 42}

Note that we are fine-tuning an already trained NLI system so we kept the number of epochs and learning-rate low. The rest of state-of-the-art systems were trained using the hyperparameters reported by the authors.

C TACRED templates

This section describes the templates used in the TACRED experiments. We performed all the experiments using the templates showed in Tables 1 (for PERSON relations) and 2 (for ORGANIZATION relations). These templates were manually created based on the TAC KBP Slot Descriptions[15] (annotation guidelines). Besides the templates, we also report the valid argument types that are accepted on each relation.

[15] https://tac.nist.gov/2014/KBP/ColdStart/guidelines/TAC_KBP_2014_Slot_Descriptions_V1.4.pdf

Relation	Templates	Valid argument types
per:alternate_names	{subj} is also known as {obj}	PERSON, MISC
per:date_of_birth	{subj}'s birthday is on {obj}	DATE
	{subj} was born on {obj}	
per:age	{subj} is {obj} years old	NUMBER, DURATION
per:country_of_birth	{subj} was born in {obj}	COUNTRY
per:stateorprovince_of_birth	{subj} was born in {obj}	STATE_OR_PROVINCE
per:city_of_birth	{subj} was born in {obj}	CITY, LOCATION
per:origin	{obj} is the nationality of {subj}	NATIONALITY, COUNTRY, LOCATION
per:date_of_death	{subj} died in {obj}	DATE
per:country_of_death	{subj} died in {obj}	COUNTRY
per:stateorprovince_of_death	{subj} died in {obj}	STATE_OR_PROVINCE
per:city_of_death	{subj} died in {obj}	CITY, LOCATION
per:cause_of_death	{obj} is the cause of {subj}'s death	CAUSE_OF_DEATH
per:countries_of_residence	{subj} lives in {obj}	COUNTRY, NATIONALITY
	{subj} has a legal order to stay in {obj}	
per:statesorprovinces_of_residence	{subj} lives in {obj}	STATE_OR_PROVINCE
	{subj} has a legal order to stay in {obj}	
per:city_of_residence	{subj} lives in {obj}	CITY, LOCATION
	{subj} has a legal order to stay in {obj}	
per:schools_attended	{subj} studied in {obj}	ORGANIZATION
	{subj} graduated from {obj}	
per:title	{subj} is a {obj}	TITLE
per:employee_of	{subj} is a member of {obj}	ORGANIZATION
per:religion	{subj} belongs to {obj}	RELIGION
	{obj} is the religion of {subj}	
	{subj} believe in {obj}	
per:spouse	{subj} is the spouse of {obj}	PERSON
	{subj} is the wife of {obj}	
	{subj} is the husband of {obj}	
per:children	{subj} is the parent of {obj}	PERSON
	{subj} is the mother of {obj}	
	{subj} is the father of {obj}	
	{obj} is the son of {subj}	
	{obj} is the daughter of {subj}	
per:parents	{obj} is the parent of {subj}	PERSON
	{obj} is the mother of {subj}	
	{obj} is the father of {subj}	
	{subj} is the son of {obj}	
	{subj} is the daughter of {obj}	
per:siblings	{subj} and {obj} are siblings	PERSON
	{subj} is brother of {obj}	
	{subj} is sister of {obj}	
per:other_family	{subj} and {obj} are family	PERSON
	{subj} is a brother in law of {obj}	
	{subj} is a sister in law of {obj}	
	{subj} is the cousin of {obj}	
	{subj} is the uncle of {obj}	
	{subj} is the aunt of {obj}	
	{subj} is the grandparent of {obj}	
	{subj} is the grandmother of {obj}	
	{subj} is the grandson of {obj}	
	{subj} is the granddaughter of {obj}	
per:charges	{subj} was convicted of {obj}	CRIMINAL_CHARGE
	{obj} are the charges of {subj}	

Table 1: Templates and valid arguments for PERSON relations.

Relation	Templates	Valid argument types
org:alternate_names	{subj} is also known as {obj}	ORGANIZATION, MISC
org:political/religious_affiliation	{subj} has political affiliation with {obj} {subj} has religious affiliation with {obj}	RELIGION, IDEOLOGY

org:top_memberts/employees	{obj} is a high level member of {subj}
	{obj} is chairman of {subj}
	{obj} is president of {subj}
	{obj} is director of {subj}
		PERSON

org:number_of_employees/members	{subj} employs nearly {obj} people {subj} has about {obj} employees	NUMBER

org:members	{obj} is member of {subj}
	{obj} joined {subj}
		ORGANIZATION, COUNTRY

org:subsidiaries	{obj} is a subsidiary of {subj} {obj} is a branch of {subj}	ORGANIZATION, LOCATION
org:parents	{subj} is a subsidiary of {obj} {subj} is a branch of {obj}	ORGANIZATION, COUNTRY

org:founded_by	{subj} was founded by {obj}
	{obj} founded {subj}
		PERSON

org:founded	{subj} was founded in {obj}
	{subj} was formed in {obj}
		DATE

org:dissolved	{subj} existed until {obj}
	{subj} disbanded in {obj}
	{subj} dissolved in {obj}
		DATE

org:country_of_headquarters	{subj} has its headquarters in {obj} {subj} is located in {obj}	COUNTRY
org:stateorprovince_of_headquarters	{subj} has its headquarters in {obj} {subj} is located in {obj}	STATE_OR_PROVINCE
org:city_of_headquarters	{subj} has its headquarters in {obj} {subj} is located in {obj}	CITY, LOCATION
org:shareholders	{obj} holds shares in {subj}	ORGANIZATION, PERSON
org:website	{obj} is the URL of {subj} {obj} is the website of {subj}	URL

Table 2: Templates and valid arguments for ORGANIZATION relations.

Textual Entailment for Event Argument Extraction: Zero- and Few-Shot with Multi-Source Learning

Oscar Sainz[1], **Itziar Gonzalez-Dios**[1],
Oier Lopez de Lacalle[1], **Bonan Min**[2], and **Eneko Agirre**[1]

[1]HiTZ Basque Center for Language Technologies - Ixa NLP Group
University of the Basque Country UPV/EHU
[2]Raytheon BBN Technologies
oscar.sainz@ehu.eus

Abstract

Recent work has shown that NLP tasks such as Relation Extraction (RE) can be recasted as Textual Entailment tasks using verbalizations, with strong performance in zero-shot and few-shot settings thanks to pre-trained entailment models. The fact that relations in current RE datasets are easily verbalized casts doubts on whether entailment would be effective in more complex tasks. In this work we show that entailment is also effective in Event Argument Extraction (EAE), reducing the need of manual annotation to 50% and 20% in ACE and WikiEvents respectively, while achieving the same performance as with full training. More importantly, we show that recasting EAE as entailment alleviates the dependency on schemas, which has been a roadblock for transferring annotations between domains. Thanks to the entailment, the multi-source transfer between ACE and WikiEvents further reduces annotation down to 10% and 5% (respectively) of the full training without transfer. Our analysis shows that the key to good results is the use of several entailment datasets to pre-train the entailment model. Similar to previous approaches, our method requires a small amount of effort for manual verbalization: only less than 15 minutes per event argument type is needed, and comparable results can be achieved with users with different level of expertise.

1 Introduction

Building Information Extraction (IE) systems for real-world applications is very costly and has suffered from data-scarcity problems, due in part to the expertise and time required to annotate training data at a large scale with sufficient consistency, but also due to poor transfer between domains: IE annotations depend on the schema used in each domain, and moving to new domains requires new schemas, new annotation guidelines and the manual annotation of new data. In many cases, there is some information overlap between schemas, but performing transfer learning to leverage such overlap (i.e. learning from **multiple sources**) can be difficult: it often requires manually mapping labels between schemas, which is typically brittle, cumbersome and requires costly domain expertise (Kalfoglou and Schorlemmer, 2003).

In order to save annotation effort, recent work recasts IE tasks as Textual Entailment tasks (White et al., 2017; Poliak et al., 2018a; Levy et al., 2017; Sainz et al., 2021). For instance, Sainz et al. (2021) manually verbalize each relation type in the Relation Extraction (RE) dataset TACRED (Zhang et al., 2017) to generate hypotheses for each test example, and then apply an entailment model to output the relation type of the hypothesis with highest entailment probability. The entailment model is typically based on large language models pre-trained on entailment datasets such as MNLI (Williams et al., 2018). The approach obtains very strong results on zero-shot and few-shot scenarios, but we note that TACRED contains relations between two entities that are easily verbalizable,[1] casting doubts on whether entailment would be effective in more complex IE tasks. Event Argument Extraction (EAE) involves more complex contexts, higher ambiguity in the words that trigger events, and depends on the event type in addition to the relation (see Figure 1).

In this work, we present the first system for EAE that addresses the task as an entailment problem. We empirically show the robustness of the method on the zero-shot, few-shot and full training regimes, obtaining state-of-the-art results on ACE (Walker et al., 2006) and WikiEvents (Li et al., 2021b). In addition, we make the following contributions: (1) We show that our method reduces schema dependency, as it improves the performance on the WikiEvents results using additional ACE training data and vice versa with no extra manual work. (2)

[1] For instance, PER:DATE_OF_BIRTH can be verbalized as `{subj}'s birthday is on {obj}` in which subj and obj refers to the two text mentions involved in the relation.

Findings of the Association for Computational Linguistics: NAACL 2022, pages 2439 - 2455
July 10-15, 2022 ©2022 Association for Computational Linguistics

Ablation results show that training with several NLI datasets is significantly better than just using MNLI. (3) Our analysis of the manual work required for writing templates and annotating arguments sheds light in the sweet spot for future applications, and shows that template writing does not require much domain expertise as shown by the results using an independent novice template writer. We make the code, templates and models publicly available.[2]

2 Related Work

Textual Entailment Given a textual premise and a hypothesis, the task is to decide whether the premise entails or contradicts (or is neutral to) the hypothesis (Dagan et al., 2006). The current state-of-the-art uses large pre-trained Language Models (LM) (Lan et al., 2020; Liu et al., 2019; Conneau et al., 2020; Lewis et al., 2020; He et al., 2021) fine-tuned on manually annotated datasets such as SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), FEVER (Thorne et al., 2018) or ANLI (Nie et al., 2020). The task is also known as Natural Language Inference (NLI).

Prompt and Pivot task based learning has emerged as a candidate solution for data-scarcity problems (Le Scao and Rush, 2021; Min et al., 2021; Liu et al., 2021a). The use of discrete (Gao et al., 2021; Schick and Schütze, 2021a,b,c) or continuous (Liu et al., 2021b) prompts allowed language models to perform significantly better on many text classification tasks. Closely related to our approach, several works make use of a high-resource supervised task such as Question Answering or entailment as pivot tasks (Yin et al., 2019, 2020; Wang et al., 2021; Sainz and Rigau, 2021; McCann et al., 2018). In the case of entailment, Dagan et al. (2006) converted QA data to entailment manually and Demszky et al. (2018) did it automatically. Other semantic tasks such as Named Entity Recognition, Relation Extraction and Semantic Role Labelling have also been reformulated as entailment by automatically converting data into the entailment format (White et al., 2017; Poliak et al., 2018a; Levy et al., 2017; Sainz et al., 2021).

Multi-task learning reformulates multiple tasks to a single and common task via prompting large pre-trained language models, leveraging multiple data sources to improve each task of interest. Such approaches have shown improvements in supervised (Subramanian et al., 2018; Raffel et al., 2020; Aribandi et al., 2022) and zero-shot scenarios (Sanh et al., 2022; Wei et al., 2021a). While using the language modelling task as a pivot shows strong performance with very large language models, it is not clear that smaller models can benefit from this strategy in the same way. Wei et al. (2021a) and Mishra et al. (2022) obtained contradictory results. In a similar way, Question Answering has been proposed as a pivot task for multi-task learning but without promising results (McCann et al., 2018). In this work, we explore multi-source learning, where datasets from different or similar tasks are used to build a model for the target task.

Event Argument Extraction is a sub-task of Event Extraction. The goal is to identify arguments or fillers for a specific slot (a.k.a., role) in an event template. This task has been largely explored on the Message Understanding Conference (MUC, Grishman and Sundheim (1996)) and later on Automatic Content Evaluation (ACE). ACE focused mainly on sentence level evaluation due to the difficulty of the task at the time. Recently, new benchmarks such as RAMS (Ebner et al., 2020) and WikiEvents have emerged with the aim of addressing document level information extraction similar to MUC. However, most of the interest is still focused on the sentence level.

EAE has been recently addressed by end-to-end event extraction models (Wadden et al., 2019; Lin et al., 2020; Li et al., 2021a), instead of treating it as an independent task (Du and Cardie, 2020a), as we do, or as a subtask in a pipeline (Lyu et al., 2021). Lately, with the recent paradigm shift **to prompt design learning** (Min et al., 2021), several works reformulated the task as a Question Answering problem (Li et al., 2020; Feng et al., 2020; Du and Cardie, 2020b; Liu et al., 2020; Wei et al., 2021b; Lyu et al., 2021; Sulem et al., 2022) or as a Constrained Text Generation problem (Chen et al., 2020; Du et al., 2021; Li et al., 2021b) using predefined prompts, questions or templates. We instead reformulate the task as a textual entailment problem.

3 Approach

In order to cast EAE as an entailment task, we verbalize event argument instances using a set of intuitive and linguistically motivated templates to capture the event argument roles, and then per-

[2] https://github.com/osainz59/Ask2Transformers

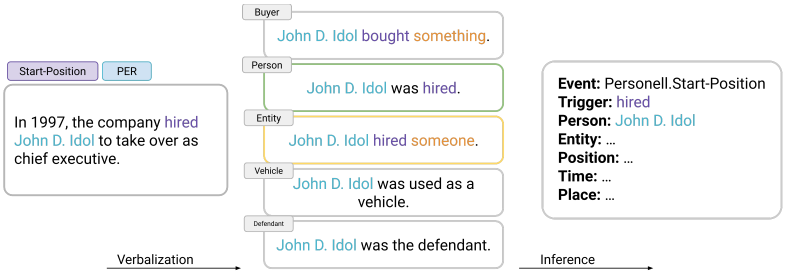

Start-Position
PER
In 1997, the company hired John D. Idol to take over as chief executive.
Verbalization
Buyer
John D. Idol bought something.
Person
John D. Idol was hired.
Entity
John D. Idol hired someone.
Vehicle
John D. Idol was used as a vehicle.
Defendant
John D. Idol was the defendant.
Inference
Event: Personell.Start-Position
Trigger: hired
Person: John D. Idol
Entity: ...
Position: ...
Time: ...
Place: ...

Figure 1: Entailment-based Event Argument Extraction. On the left, input information: the context, the event trigger (*hired*) and the argument candidate (*John D. Idol*), alongside the types of both. On the middle, some hypothesis verbalized using the templates: the green box is entailed, the yellow box matches the type constraint but it is not entailed, and the rest do not satisfy type constraints. On the right, the output with the inferred role (Person).

form inferences with entailment models. The entailment model can be additionally trained with EAE training data converted into the entailment format, similar to Sainz et al. (2021). Figure 1 shows the general workflow of the method. First, the possible roles are verbalized by means of predefined templates and the input, which comprises the context, trigger and argument candidate. Then, an entailment model is used to generate the entailment probability for each verbalization. To predict the role, the most probable hypothesis (verbalization) is chosen among the roles that satisfy the event-entity[3] constraints. A more detailed description of each component follows.

Label verbalization is attained using templates that combine the information of the instance and express a specific label. Different role verbalizations are shown in Figure 1. A verbalization is generated using templates that have been manually written based on the task guidelines of each dataset. The templates involve the candidate argument, and optionally the event trigger. In some cases, in order to produce a grammatical hypothesis, placeholders corresponding to the agent or theme are also introduced, which can be generic, e.g. *someone*, or dependent of the argument role, e.g. *defendant*. We defined several template types (see Table 1) to guide the creation of templates more systematically. In Section 5.1 we describe the process to create templates, and in Section 7 we analyse the differences between independent template developers and how this did not affect performance. The templates created for the ACE dataset are listed in Appendix C.

Entailment model. Given a premise and hypothesis, the model returns the probabilities of the hypothesis being entailed by, contradicted to or neutral to the premise. In principle, any model trained on the NLI task can be used.

Inference takes into account three key factors to output the role label for an argument candidate: the entailment probabilities of each verbalization, the type constraints of the specific role, and a threshold. Argument candidates which do not match the type constraints are discarded. From the rest, we return the role of the verbalized hypothesis with highest entailment probability, unless the probability is lower than the threshold, in which case we return the negative class.[4]

Training. Our entailment-based model can be applied without any training on the EAE task, in a zero-shot fashion, or, alternatively, the entailment model can be finetuned using training data from the EAE dataset. For this purpose, we convert the EAE training dataset into a NLI format, i.e we generate entailment, neutral and contradiction hypotheses heuristically from the data using the templates themselves. For each positive labeled example (a candidate that is an argument) we sample N_E entailment hypotheses using the templates that correspond to the correct label and N_N neutral hypotheses using templates from different roles. For each negative example (the candidate is not an argument of the event) we create N_C contradiction hypotheses using any template at random. N_E, N_N and N_C are considered hyperparameters of the training phase along with the hyperparameters of the neural network model such as learning-rate and

[3] In this context, entities also include values such as time or amounts.

[4] The class that represents that the argument candidate takes no part on the event.

Template type	Description	Example
{arg}	Templates with **implicit** information about the event. {arg} variable is the placeholder for the argument candidate.	The victim was {arg}.
{trg} → {arg}	Templates with **explicit** information about the event. The {trg} variable is the placeholder for the event trigger.	The {trg} occurred in {arg}.
{canonical(trg)} → {arg}	Templates with predefined canonical values for the {trg} variable.	{arg} was jailed.
{canonical(trg)}, placeholder → {arg}	Templates that makes use of agent or patient dummy placeholders in order to produce grammatical sentences.	The {arg} inspected something.

Table 1: The four main template categories used to create the role verbalizations.

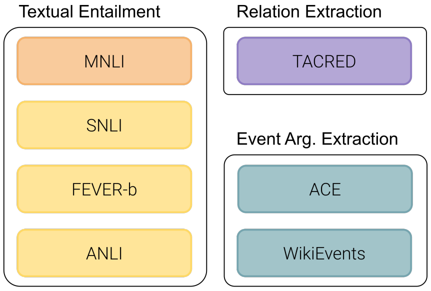


Figure 2: Datasets used by task category.

batch-size. In order to create challenging training examples for the negative class, we propose to use **constrained sampling**, based on the trigger-entity type constraints, where we create negative examples from candidates that satisfy the constraints. Preliminary experiments showed slight improvements with respect to regular sampling.

4 Entailment for Multi-source Learning

We hypothesize that two similar IE tasks can benefit from each other even if they do not share the same schema or domain. Although this hypothesis is very intuitive and it has been demonstrated on several works for tasks other than IE (see Multi-task learning on Section 2), actual IE models are limited by schema dependency, which makes it almost impossible to learn from datasets annotated with different IE schemas. One option is to perform a manual mapping between schemas, which is costly and often inaccurate (Kalfoglou and Schorlemmer, 2003). Our approach instead is domain and schema agnostic, and therefore allows to learning from multiple sources seamlessly. Given that the sources are recast into a single format in a common entailment formulation, it suffices to fine-tune the model in sequence across the sources.

To check our hypothesis we split tasks according to the following criteria: (1) IE sources like Relation Extraction that are different from EAE (e.g. TACRED), and (2) EAE sources using different schemas (e.g. WikiEvents and ACE). Figure 2 summarizes the tasks and datasets used in this work, including the four natural language understanding datasets.

5 Experimental Setup

In this section, we describe the methodology for template development, evaluation setting, the baselines used in our experiments, and the computation infrastructure specifications.

5.1 Methodology for verbalization

The templates used to generate the verbalizations were created based on the annotation guidelines of each dataset. During the creation, the template developers had access to the guidelines that describe each of the roles (which can include one or two examples) and a NLI model that the developer could use to verify whether the generated verbalizations of these examples were entailed by the model. The developer was allowed a maximum of 15 minutes per role, and spent 5 and 12 hours[5] to create the templates for ACE and WikiEvents respectively.

5.2 Evaluation

Datasets. We carried out our evaluation on two different EAE datasets: ACE (Walker et al., 2006) and WikiEvents (Li et al., 2021b). The ACE2005 dataset is a sentence-level Event Extraction dataset that contains entities, relations, event-triggers and arguments annotations on English, Chinese and

[5] Given that there is a total of 22 and 59 role types respectively, this is equivalent to an average of 13 and 12 minutes per role.

	ACE		WikiEvents	
Train split	# Pos	Total	# Pos	Total
0%	-	-	-	-
1%	2.05	173	0.86	195
5%	11.36	843	4.09	966
10%	23.86	1736	8.26	1903
20%	45.00	3302	15.84	3578
100%	220.86	16502	79.68	18532

Table 2: Mean examples per role (pos) and total number of examples (positive and negative) across different training data splits and datasets.

Arabic texts. We worked only on the English EAE task. The WikiEvents dataset is instead more focused on document-level argument extraction task. Although the last is intended to be use as a document-level benchmark we focused on the sentence-level extraction[6] for two reasons: to maintain consistency with ACE dataset and because the nearest occurrence of the arguments are inside the sentence of the event trigger in almost all examples. For both ACE and WikiEvents, we split the training data into different amounts (0%, 1%, 5%, 10%, 20% and 100%) following Liu et al. (2020) to also evaluate our system on extreme data scarcity scenarios. Table 2 shows the amount of examples per split. The total amount refers to the addition of all positives and negatives trigger-candidate pairs.

Metrics. We have used the standard F1-Score, which is a common metric on IE tasks. Along with that, we propose the use of the Area Under the Curve (AUC) for better model comparison across all scenarios. The reported AUC scores are computed with all splits for the main results and just with 0%, 5% and 100% for the multi-source results, and therefore, they are not comparable.

5.3 Baselines and Models

Baselines. Our main point of comparison is our re-implementation of EM (Baldini Soares et al., 2019), as we can run it on the same few-shot splits as our system and allow for head-to-head comparison. EM is a state-of-the-art (Zhou and Chen, 2021) model that uses RoBERTa$_{\text{LARGE}}$ as a backbone. In addition we also report results of the state-of-the-art models that have been run on our same experimental setup, having access to gold event-trigger and entity annotations. On ACE, we report the results of BERTEE and RCEE_ER, both reported at (Liu et al., 2020), which correspond to a BERT (Devlin et al., 2019) based baseline and a QA based pivot approach that leverages SQuAD (Rajpurkar et al., 2016) data. Unfortunately the data splits used by (Liu et al., 2020) are not available[7] and thus, only the results for zero-shot (i.e. 0% training data) and full training (i.e. 100% training data) are directly comparable. Regarding WikiEvents Gen-Arg (Li et al., 2021b) uses gold triggers, but not gold entity information, so we decided to report Coref-F1[8] which refers to the F1-Score of predicting at least one of the gold entity coreferential chain as argument.

NLI models used in this work are based on the RoBERTa$_{\text{large}}$ (Liu et al., 2019) checkpoint, and are available via HuggingFace Transformer's model repository (Wolf et al., 2020). The main results use a model trained on all MNLI, SNLI, FEVER and ANLI, and in the analysis we also report the results of a model using just MNLI (see Appendix A for more information, including hyperparameters used).

5.4 Infrastructure

All the experiments were done in a **single** RTX 2080ti (11Gb) with a 250W power consumption. The average training times are:[9] 0.36h/epoch for ACE, 0.52h/epoch for WikiEvents and 2.86 h/epoch for TACRED. In total, 464.56 hours (154.86 if only a single run is done) of computation time are required to reproduce **all** the experiments, that in our setting corresponds to 21.36 kgCO$_2$eq carbon footprint[10] (roughly equivalent to the CO$_2$ emitted by 88.2 km driven by an average car).

6 Results

Main results. Table 3 reports our NLI system, including the median F1-Score and the standard deviation across 3 different runs of our implementations NLI and EM. On ACE our system is best on all comparable results and overall as shown by the AUC score. On the case of WikiEvents, our

[6] We consider as model prediction errors the arguments that are outside the sentence, to be consistent with other systems evaluation.

[7] Personal communication.

[8] We used this to alleviate the noise introduced by not using the gold entity annotations, and therefore, make the comparison more fair.

[9] The time required for training the model depends linearly with the sampling rates of entailment, neutral and contradiction examples.

[10] Estimation based on mlco2.github.io/impact/

				ACE			
Model	0%	1%	5%	10%	20%	100%	AUC
BERTEE	-	*2.20	*10.5	*19.3	*28.6	64.7	*40.73
EM	-	4.58 ±1.55	37.5 ±2.98	50.9 ±0.96	58.7 ±1.9	72.1 ±0.65	60.87
RCEE_ER	37.0	*49.8	*59.9	*65.1	*67.6	70.1	*67.47
NLI	**40.6**	**45.4** ±0.16	**57.1** ±0.93	**64.6** ±1.12	**69.8** ±0.58	**74.6** ±0.88	**70.00**
				WikiEvents			
Model	0%	1%	5%	10%	20%	100%	AUC
EM	-	16.9 ±0.63	41.5 ±1.47	49.9 ±0.28	54.9 ±1.30	61.3 ±1.04	55.26
*Gen-Arg	-	2.4 ±1.66	30.5 ±4.12	48.1 ±1.42	55.7 ±1.35	65.1	56.15
NLI	**35.9**	**42.6** ±1.36	**52.2** ±1.40	**59.5** ±0.58	**65.4** ±0.62	**69.9** ±0.70	**65.45**

Table 3: Main results on different training data splits for our NLI model, EM baseline and state-of-the-art systems. * for results not directly comparable with ours. Bold for best among comparable results.

	ACE				WikiEvents			
Source	0%	5%	100%	AUC	0%	5%	100%	AUC
NLI	40.6	57.1 ±0.93	74.6 ±0.88	65.0	35.9	52.2 ±1.40	69.9 ±0.70	60.2
NLI $_{+\text{ WikiEvents}}$	**62.7**	**69.3** ±0.35	**74.9** ±0.58	**71.8**	-	-	-	-
NLI $_{+\text{ ACE}}$	-	-	-	-	**57.3**	65.2 ±0.41	**71.5** ±1.07	**68.0**
NLI $_{+\text{ RE}}$	44.5	56.3 ±0.79	73.9 ±0.05	64.4	38.2	55.0 ±1.38	69.2 ±0.59	61.3
NLI $_{+\text{ RE + WikiEvents}}$	**62.7**	65.9 ±0.30	74.0 ±0.49	69.7	-	-	-	-
NLI $_{+\text{ RE + ACE}}$	-	-	-	-	56.7	**66.4** ±0.95	69.8 ±2.68	67.8

Table 4: Multi-source learning results of the NLI model. The AUC score reported on this table is only computed with 0%, 5% and 100% points, and therefore, is not comparable with Table 3. RE is shorthand for TACRED.

system is the best in all cases. In both datasets the EM baseline is outperformed by the NLI system.

Multi-source results. Table 4 describes our multi-source learning results, where we use NLI+ to indicate systems that use additional sources for training. We report the median F1-Score across 3 runs for 0%, 5% and 100% scenarios and the corresponding AUC score on ACE and WikiEvents. The rows show the impact of transferring knowledge from the training part of different tasks (for more detailed per role analysis see Appendix B). The results show that the signal between EAE datasets (i.e. WikiEvents and ACE) is strong, yielding significant improvements in all scenarios. For instance, on zero-shot evaluation, the systems obtain the impressive scores of 62.7 and 57.3, close to 20 points of improvement.

Sequentially fine-tuning our NLI model in TACRED and then in our target task shows small improvements on low-resource scenarios (0% split for ACE, 0% and 5% splits for WikiEvents). Training on the three sources sequentially does not seem to yield further improvements.

Figure 3 shows the performance of our NLI and multi-source enhanced NLI+ systems along with the EM baseline (data from Tables 3 and 4). The curves show that our NLI+ systems only need 10% and 5% of the data (on ACE and WikiEvents, respectively) to outperform the EM baseline that uses 100% of the training data.

7 Analysis

After performing the main experiments we did some additional analysis.

The importance of using several NLI datasets. A perfect NLI model should, in theory, solve any task that is framed correctly as entailment. Of course, there is not "perfect" NLI model. In fact, current state-of-the-art NLI models tend to learn artifacts and lexical patterns (Gururangan et al., 2018; Poliak et al., 2018b; Tsuchiya, 2018; Glockner et al., 2018; Geva et al., 2019; McCoy et al.,

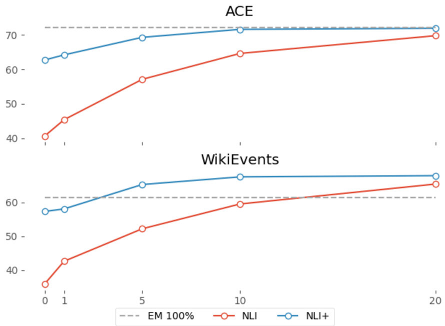


Figure 3: Comparison between the baseline EM model trained on 100% training, and our NLI and multi-source enhanced NLI+ models (NLI$_{+\text{ WikiEvents}}$ and NLI$_{+\text{ ACE}}$) with different training subsets.

Model $_{\text{source}}$	0%	5%	100%	AUC
		ACE		
NLI $_{\text{MNLI only}}$	31.4	46.0 ±0.55	62.8 ±2.83	53.6
NLI	**40.6**	**57.1** ±0.93	**74.6** ±0.88	**65.0**
		WikiEvents		
NLI $_{\text{MNLI only}}$	29.5	49.3 ±0.32	59.9 ±0.99	53.8
NLI	**35.9**	**52.2** ±1.40	**69.9** ±0.70	**60.2**
		TACRED		
NLI $_{\text{MNLI only}}$	55.6	64.1±0.20	71.0	67.2
NLI	**56.8**	**70.5**±0.62	**73.2**±0.65	**71.4**

Table 5: Ablation on NLI datasets used to-pretrain our NLI model on three datasets. NLI for our system using MNLI, FEVER, SNLI and ANLI (taken Table 3) and NLI $_{\text{MNLI only}}$ for our system when using MNLI only.

2019) instead of the task itself. Motivated by these issues, datasets like ANLI (Nie et al., 2020) were adversarially created to alleviate them. The lack of robustness of NLI models gets amplified when it comes to a cross-task evaluation. For instance, the model trained on MNLI achieves 90.2 accuracy on MNLI and 31.4, 29.5 and 55.6 F1-Score on ACE, WikiEvents and TACRED respectively (cf. Table 5). Adding FEVER, SNLI and ANLI to the training improves MNLI accuracy only 0.8 points to 91.0, but zero-shot scores on ACE, WikiEvents and TACRED improve +9.2, +6.4 and +1.2 respectively. In few-shot and full-training scenarios, the results also improve when using several NLI datasets. Our results suggest that new, more challenging NLI datasets, as well as NLI datasets automatically generated from other sources (as done in this work with WikiEvents and ACE) will yield more robust entailment models, and could further increase the performance of entailment-based EAE and IE.

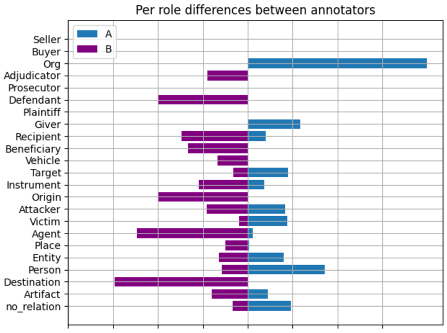


Figure 4: Recall differences between the main developer (A, right) and the linguist (B, left).

The impact of different template developers. In order to test the robustness of the templates, we enrolled a linguist with experience in NLP annotation but no prior contact with the project nor access to the original templates from the main developer. Under the same time and resource conditions, she was asked to write templates for the ACE dataset. The templates written by the main developer and the linguist vary in different ways: (1) the number of created templates per role and (2) the verbalization style, as the main developer tended to use finite and conjugated verbs while the linguist tended to use infinitives and lemmas. The templates of both are available in Appendix C.

To study the performance of the templates of each developer per role, Figure 4 shows the instances that a system correctly classified and the other system did not, and vice versa. The bars display the recall, as they are normalized by the frequencies of the roles. Missing bars on a row means that both performed the same on that role (e.g. Seller). When only a blue bar is shown (e.g. Org) it means that the main developer recovered arguments which the linguist did not, **and** there were no examples where the linguist recovered arguments that the developer did not. The same applies to situations where there is only purple bars. Roles with mixed results include examples where one or the other succeeded. As we can see, the approaches seem to be complementary, with the linguist having a higher recall with the roles that are more associated with classical semantic roles. Table 6 shows that in general, the templates of the linguist perform similarly to those of the main developer, except for 100% of the data, where the templates of the main developer were slightly better.

Developer	0%	1%	5%	10%	20%	100%
(A) Main	40.3	46.2	56.3	63.8	69.6	76.4
(B) Linguist	40.4	44.9	57.3	64.2	70.1	73.3
∆ F1	-0.1	+1.3	-1.0	-0.4	-0.5	+3.1

Table 6: Results for templates from two developers. Median F1 on the development set are reported.

Verbalizations vs. annotations Finally, we carried out an experiment to compare the time and effort requirements of annotation vs. writing the templates. To that end, the linguist re-annotated a small portion of ACE with the same information she had as she was creating the templates. That is, given the argument candidates for each event trigger in the document, she needs to decide whether the candidate was an argument and the type of the argument. She has access to the guidelines (similar to creating the templates), though she did not study them beforehand. Note also that she did the annotations **after** writing the templates, so she was already familiar with the slots. Under these conditions, she annotated 46 pairs (event trigger, potential argument candidate) in 30 minutes. Taking into account that ACE has 16.5000 such pairs, it would take approximately 180 hours to annotate ACE training part. Note that in practice, ACE requires much more time than our estimate to achieve the desired level of quality: the ACE annotation procedure involved double annotation and a second pass with a senior annotator (Doddington et al., 2004). For an analysis of the annotation procedure the interested reader is referred to Min and Grishman (2012).

Based on our estimation, 9 hours would allow an annotator to annotate 5% of the dataset which yields a 37.5 F1 (Figure 5), while 5 hours of template building yields 40.6 F1-Score in the zero-shot setting. With 18 hours 10% would be annotated and the F1-Score will be 50.9, while 5 hours of template building and 9 hours of annotations would yield 57. Figure 5 plots the performance according to manual hours on ACE, showing the huge gains provided by the initial 5 hours writing templates, plus the reuse of WikiEvents annotations. According to our experience, more hours on template building does not necessarily lead to improvements (contrary to annotation), so a **sweet spot for time investment** seems to be to firstly create templates, and then spend the remaining budget on annotating examples.

On another note, the linguist mentioned that writing templates is more natural and rewarding than annotating examples, which is more repetitive, stressful and tiresome. When writing templates, she was thinking in an abstract manner, trying to find generalizations, while she was paying attention to concrete cases when doing annotation.

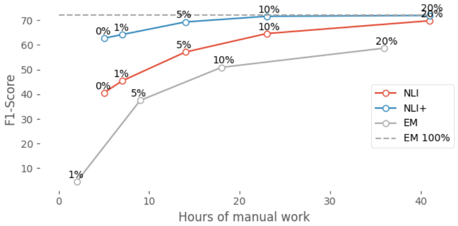


Figure 5: Performance on ACE according to our estimations of manual work in hours. We also indicate the percentage of training data used.

8 Conclusions

This paper shows the entailment-based approach for event argument extraction is extremely effective in zero-shot, few-shot and full train scenarios both on ACE and WikiEvents, outperforming previous methods. First of all, recasting EAE as an entailment task allows it to reuse annotations from different event schemas, achieving large gains when transferring annotations between ACE and WikiEvents, and also some gains in the zero-shot performance when transferring annotations from a relation extraction model such as TACRED. Secondly, we show that using additional training entailment datasets improves results significantly over just using MNLI, not only on EAE but also on TACRED. Thirdly, we show that the relatively short time spent writing manual templates is much more effective than the time spent on doing annotations, with a sweet spot where the annotation effort is split between the two, with large savings in manual labour. Lastly, we show that an independent linguist is able to write templates with comparable performance without any special training. We think that our results and analysis support the potential of entailment models for other NLP tasks.

Our work paves the way for a new paradigm for IE, where the expert defines the schema using natural language and directly runs those specifications, annotating a handful of examples in the process, and allowing for quick trial-and-error iterations. Sainz et al. (2022) propose a user interface alongside this paradigm. More generally, inference capability could be extended, acquired and applied from other tasks, in a research avenue where entailment and task performance improve in tandem.

Acknowledgements

Oscar is funded by a PhD grant from the Basque Government (PRE_2020_1_0246). This work is based upon work partially supported via the IARPA BETTER Program contract No. 2019-19051600006 (ODNI, IARPA), and by the Basque Government (IXA excellence research group IT1343-19).

References

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2022. Ext5: Towards extreme multi-task scaling for transfer learning. In *International Conference on Learning Representations*.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Yunmo Chen, Tongfei Chen, Seth Ebner, Aaron Steven White, and Benjamin Van Durme. 2020. Reading the manual: Event extraction as definition comprehension. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP*, pages 74–83, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation. In *Language Resources and Evaluation Conference (LREC)*, volume 2, pages 837–840. Lisbon.

Xinya Du and Claire Cardie. 2020a. Document-level event role filler extraction using multi-granularity contextualized encoding.

Xinya Du and Claire Cardie. 2020b. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Xinya Du, Alexander Rush, and Claire Cardie. 2021. GRIT: Generative role-filler transformers for document-level event entity extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 634–644, Online. Association for Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Rui Feng, Jie Yuan, and Chao Zhang. 2020. Probing and fine-tuning reading comprehension models for few-shot event extraction.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference* *on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Yannis Kalfoglou and Marco Schorlemmer. 2003. Ontology mapping: the state of the art. *The knowledge engineering review*, 18(1):1–31.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838, Online. Association for Computational Linguistics.

Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, Jiawei Han, and Clare Voss. 2021a. The future is not one-dimensional: Complex event schema induction by graph modeling for event prediction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5203–5215, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021b. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021b. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Qing Lyu, Hongming Zhang, Elior Sulem, and Dan Roth. 2021. Zero-shot event extraction via transfer learning: Challenges and insights. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 322–332, Online. Association for Computational Linguistics.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Bonan Min and Ralph Grishman. 2012. Compensating for Annotation Errors in Training a Relation Extractor. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 194–203.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018a. Towards a unified natural language inference framework to evaluate sentence representations. *CoRR*, abs/1804.08207.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018b. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oscar Sainz, Haoling Qiu, Oier Lopez de Lacalle, Agirre Eneko, and Bonan Min. 2022. ZS4IE: A toolkit for zero-shot information extraction with simple verbalizations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, Online and Seattle, Washington. Association for Computational Linguistics.

Oscar Sainz and German Rigau. 2021. Ask2Transformers: Zero-shot domain labelling with pretrained language models. In *Proceedings of the 11th Global Wordnet Conference*, pages 44–52, University of South Africa (UNISA). Global Wordnet Association.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021c. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations*.

Elior Sulem, Jamaal Hay, and Dan Roth. 2022. Yes, No or IDK: The challenge of unanswerable Yes/No questions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Online and Seattle, Washington. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021a. Finetuned language models are zero-shot learners.

Kaiwen Wei, Xian Sun, Zequn Zhang, Jingyuan Zhang, Guo Zhi, and Li Jin. 2021b. Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4672–4682, Online. Association for Computational Linguistics.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. 2020. Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8229–8239, Online. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2021. An improved baseline for sentence-level relation extraction.

Hyperparameter	EM	NLI	NLI$_{\text{MNLI only}}$
N$_E$ / N$_N$ / N$_C$	-	2 / 5 / 5	2 / 5 / 5
Batch size	32		
Learning rate	1×10^{-5}	4×10^{-6}	1×10^{-5}
Seeds	{0, 24, 42}		
Epochs	25 (*50)		
Weight decay	0.01		

Table 7: Hyperparameters of the trained systems. * indicates the difference between full-train and few-shot scenarios.

A Hyperparameters

On this section we describe the hyperparameters we have used on our experiments. All the hyperparameters optimized on this work were optimized for the 100% split with the batch-size fixed to 32, and used on the rest. The Table 7 describes the hyperparameters used on EM, NLI and NLI$_{\text{MNLI only}}$ variants, for the NLI+ the same hyperparameters as NLI were used. We have found that the same exact hyperparameters were the best on ACE, WikiEvents and TACRED datasets. For the future, we plan to test new hyperparameter sets that uses bigger batch-sizes, as recent works (Aribandi et al., 2022) suggest to be optimal for multi-task and -source learning experiments.

The pre-trained NLI models used on this work can be downloaded from the HuggingFace Models repository: NLI $_{\text{MNLI only}}$ (`roberta-large-mnli`) and NLI (`ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli`).

The fine-tuned models derived from this work will be uploaded to HuggingFace Models repository. Check the GitHub repository for updated information.

B Multi-task in-depth analysis

The Figure 6 shows the per role absolute improvement obtained by training on different tasks over the 0% NLI system. Overall, we can see that training on ACE or WikiEvents improves almost all the roles and training on TACRED improves some and some others do not. A result that was unexpected is that there are few roles on WikiEvents that after training on WikiEvents become worse in contrary to training on ACE. This could be explained by the differences among the frequency distributions that the train, development and test sets of WikiEvents has. Moreover, there are some roles on WikiEvents that decreases in all training scenarios, this suggests us that sequential fine-tuning might be not the best option for this type of multi-source learning and therefore further ways should be explored.

C ACE templates from both developers

The next table contains the templates written by both developers for the ACE arguments. We follow the notation introduced in Section 5.1. In addition, we also consider information from the event, such as the type on different granularity levels, including `{trg_type}` for the trigger type (e.g. *Movement* from *Movement.Transport*) and `{trg_subtype}` for the subtype of the trigger, e.g. *Transport* from *Movement.Transport*).

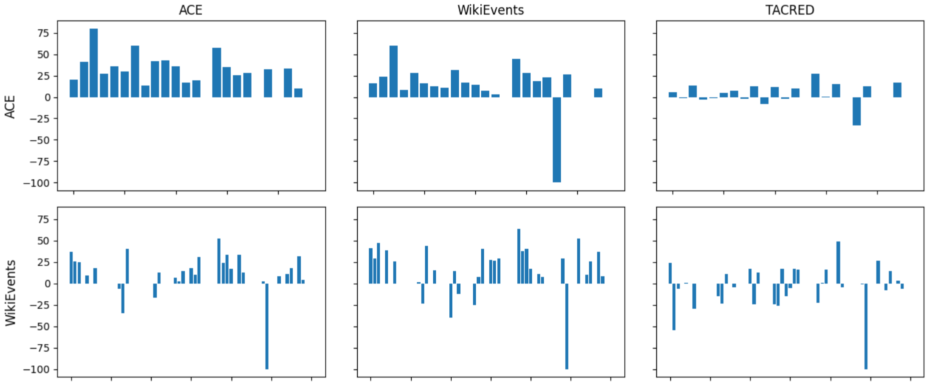


Figure 6: Absolute improvements over the NLI baseline using different tasks and sources. Rows indicates the testing data and columns the training data. Each bar indicates the F1-Score difference between the trained NLI system vs 0% NLI for a specific role.

Role	Main developer	Linguist
Adjudicator	{arg} tried the defendant. {arg} convicted the defendant. {arg} released the defendant. {arg} sentenced the defendant. {arg} acquitted the defendant.	{arg} convict someone. {arg} sentence someone. {arg} judge someone. {arg} fine someone. {arg} indict someone.
Agent	{arg} {trg} a person or organization.	{arg} do something. {arg} select something. {arg} carry out something. {arg} create something. {arg} give something.
Artifact	Someone {trg} the {arg}. Someone moved {arg}. Someone bought {arg}. Someone sold {arg}.	{arg} be an object. {arg} be a weapon.
Attacker	{arg} {trg} a person or organization.	{arg} assail someone. {arg} aggress someone. {arg} assault someone.
Beneficiary	The buyer bought to {arg} something.	{arg} get something . {arg} be beneficiary. {arg} benefit from something. {arg} obtain something.
Buyer	{arg} bought something.	{arg} buy something. {arg} possess something. {arg} own something.
Defendant	{arg} was the defendant.	{arg} be accused of something. {arg} be accused of a crime. {arg} be judged.
Destination	Someone {trg_subtype} to {arg}.	{trg_type} go to {arg}. {trg_type} finish in {arg}. {trg_type} move to {arg}. {arg} be a place. {arg} be a location.
Entity	{arg} attended the demonstration. {arg} met someone. {arg} fired someone. {arg} voted in the elections. {arg} released the defendant. {arg} was ordered to pay.	{arg} select something. {arg} carry out something. {arg} do something. {arg} create something. {arg} give something.
Giver	{arg} gave something to someone.	{arg} give something.
Instrument	Someone {trg_subtype} with {arg}.	{arg} be artifact. {arg} be object. {arg} be device. {arg} cause harm.
Org	{arg} organization declared bankruptcy.	{arg} be in bankruptcy.

(continued on the next page)

Role	Main developer	Linguist
	{arg} organization was dissolved. {arg} organization was merged. {arg} organization was launched.	{arg} be ended. {arg} be merged. {arg} be created. {arg} be company. {arg} be organization.
Origin	Someone {trg_subtype} from {arg}.	{arg} change location. {arg} be location. {trg_type} start in {arg}. {trg_type} move from {arg} .
Person	{arg} was {trg}.	{arg} be person. {arg} be living entity. {arg} be born. {arg} get married. {arg} be married. {arg} divorce. {arg}'s marriage ended. {arg} be hired. {arg} start a job. {arg} be fired. {arg} end a job. {arg} be nominated. {arg} be elected. {arg} be arrested. {arg} be jailed. {arg} be imprisoned. {arg} be released. {arg} be paroled. {arg} be executed. {arg} be extradited.
Place	{trg} occurred in {arg}.	{arg} be a place. {arg} be a location. {arg} be a placement.
Plaintiff	{arg} filed suit against someone.	{arg} bring a lawsuit against someone. {arg} bring a lawsuit against something. {arg} sue someone. {arg} sue something.
Prosecutor	{arg} indicted the defendant. {arg} charged the defendant.	{arg} prosecute. {arg} take somebody to court for a crime.
Recipient	{arg} received money from someone.	{arg} receive something. {arg} get something. {arg} get money.
Seller	{arg} sold something.	{arg} sell something.
Target	{arg} was {trg_subtype}.	{arg} be attacked. {trg_type}'s target be {arg}.
Vehicle	{arg} was used as a vehicle.	{arg} be a transport.

(continued on the next page)

Role	Main developer	Linguist
		{arg} be a vehicle. {arg} serve to move. {arg} serve to change location. {arg} serves as a means of transportation.
Victim	{arg} was {trg}.	{arg} be victim. {arg} be injured. {arg} be killed. {arg} be harmed. {arg} have a dead. {arg} have a tragedy.

The templates written by both developers for ACE.

ZS4IE: A toolkit for Zero-Shot Information Extraction with simple Verbalizations

Oscar Sainz[1,*], **Haoling Qiu**[2,*],
Oier Lopez de Lacalle[1], **Eneko Agirre**[1], and **Bonan Min**[2]

[1]HiTZ Basque Center for Language Technologies - Ixa NLP Group
University of the Basque Country (UPV/EHU)
[2]Raytheon BBN Technologies
oscar.sainz@ehu.eus, haoling.qiu@raytheon.com

Abstract

The current workflow for Information Extraction (IE) analysts involves the definition of the entities/relations of interest and a training corpus with annotated examples. In this demonstration we introduce a new workflow where the analyst directly verbalizes the entities/relations, which are then used by a Textual Entailment model to perform zero-shot IE. We present the design and implementation of a toolkit with a user interface, as well as experiments on four IE tasks that show that the system achieves very good performance at zero-shot learning using only 5–15 minutes per type of a user's effort. Our demonstration system is open-sourced at `https://github.com/BBN-E/ZS4IE`. A demonstration video is available at `https://vimeo.com/676138340`.

1 Introduction

Information Extraction (IE) systems are very costly to build. The current **define-then-annotate-and-train** workflow uses supervised machine learning, where the analyst first defines the schema with the entities and relations of interest and then builds a training corpus with annotated examples. Unfortunately, each new domain and schema requires starting from scratch, as there is very little transfer between domains.

We present an alternative **verbalize-while-defining workflow** where the analyst defines the schema interactively in a user interface using natural language verbalizations of the target entity and relation types. Figure 1 shows sample verbalization templates for a simple schema involving an employee relation and a passing away event, as well as a sample output annotated with the schema. The annotation of the EMPLOYEEOF relation requires performing Named Entity Recognition (NER) (Tjong Kim Sang and De Meulder, 2003) and Relation Extraction (RE) (Zhang et al., 2017), while annotating the LIFE.DIE event involves NER, Event Extraction (EE), and Event Argument Extraction (EAE) (Walker et al., 2006). Our toolkit is able to perform those four IE tasks using a single user interface, allowing the analyst to easily model and test the schema without the need to annotate examples.

Our toolkit leans on recent work which has successfully recast several IE tasks as Textual Entailment (TE) tasks (White et al., 2017; Poliak et al., 2018; Levy et al., 2017; Sainz et al., 2021). For instance, Sainz et al. (2021) model relation types between entity pairs using type-specific verbalization templates that describe the relation, generates a verbalization (hypothesis) automatically using those templates and then uses a pre-trained TE model to predict if the premise (the sentence where the pair appears) entails the hypothesis, therefore leading to a prediction of the relation or "no relation".

In this paper we thus present ZS4IE, a toolkit for zero-shot IE. We show that the four mainstream IE tasks mentioned above can be reformulated as TE problems, and that it is possible to achieve strong zero-shot performances leveraging pre-trained TE models and a small amount of templates curated by the user. Our toolkit allows a novice user to curate templates for each new types of entities, relations, events, and event argument roles, and validate their effectiveness online over any example. We also present strong results on widely used datasets with only 5-15 minutes per type of a user's effort.

2 Related Work

Textual Entailment has been shown to be a reasonable proxy for classification tasks like topic or sentiment analysis (Yin et al., 2019; Sainz and Rigau, 2021; Zhong et al., 2021). To reformulate a classification problem as TE, it often starts with defining templates to describe each class label, leading to a natural language text (a "verbalization" of a hypothesis) for each possible label. Inference is

[*] Denotes equal contribution.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations, pages 27 - 38
July 10-15, 2022 ©2022 Association for Computational Linguistics

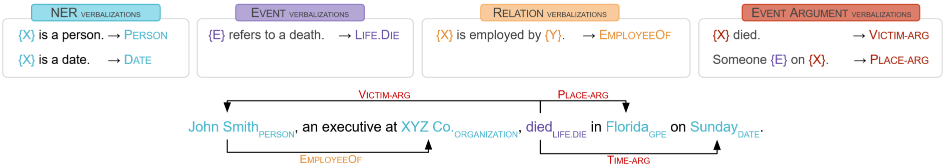

NER verbalizations
{X} is a person. → Person
{X} is a date. → Date
Event verbalizations
{E} refers to a death. → Life.Die
Relation verbalizations
{X} is employed by {Y}. → EmployeeOf
Event Argument verbalizations
{X} died. → Victim-Arg
Someone {E} on {X}. → Place-Arg
Victim-Arg
Place-Arg
John Smith PERSON, an executive at XYZ Co. ORGANIZATION, died LIFE.DIE in Florida GPE on Sunday DATE.
EmployeeOf
Time-Arg

Figure 1: Verbalization templates for a sample schema involving four tasks (from left to right, NER, EE, RE, EAE), with example output (bottom). The schema contains a EmployeeOf relation between Person and Organization entities and a Life.Die event with three argument types (Victim, Place and Time) and Person, Date and GPE entities as fillers. Due to space constraints, at most two verbalizations per task shown.

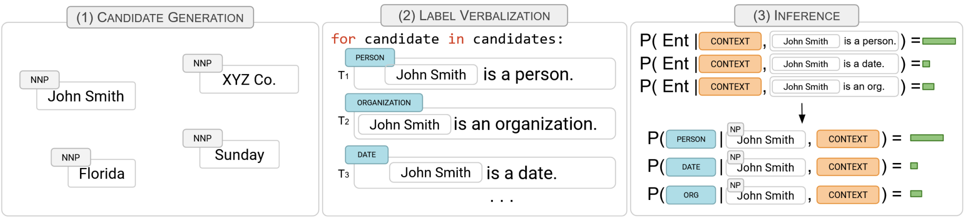

(1) Candidate Generation
NNP
John Smith
NNP
XYZ Co.
NNP
Florida
NNP
Sunday
(2) Label Verbalization
for candidate in candidates:
PERSON
T1
John Smith is a person.
ORGANIZATION
T2
John Smith is an organization.
DATE
T3
John Smith is a date.
. . .
(3) Inference
P(Ent | CONTEXT , John Smith is a person.) =
P(Ent | CONTEXT , John Smith is a date.) =
P(Ent | CONTEXT , John Smith is an org.) =
P(PERSON | NP John Smith , CONTEXT) =
P(DATE | NP John Smith , CONTEXT) =
P(ORG | NP John Smith , CONTEXT) =

Figure 2: Three steps for entailment-based NER. The steps for the other IE tasks is analogous.

performed by selecting the most probable candidate hypothesis entailing the premise. TE is usually implemented with pre-trained language model fine-tuned on TE datasets, such as MNLI (Williams et al., 2018), SNLI (Bowman et al., 2015), FEVER (Thorne et al., 2018), ANLI (Nie et al., 2020) or XNLI (Conneau et al., 2018). The results on classification have been particularly strong for zero-shot and few-shot learning, with Wang et al. (2021b) hypothesizing that entailment is a true language understanding task, where a model that performs entailment well is likely to succeed on similarly-framed tasks.

Sainz et al. (2021) reformulated relation extraction as a TE task surpassing the state-of-the-art in zero- and few-short learning. A similar approach was previously explored by Obamuyide and Vlachos (2018), using TE models that are not based on pre-trained language models. Similar to TE, (Clark et al., 2019) performs yes/no Question Answering, in which a model is asked about the veracity of some fact given a passage. Lyu et al. (2021) recast the zero-shot event extraction as a TE task, using TE model to check whether a piece of text is about a type of event. Lastly, Sainz et al. (2022) showed that TE allows to leverage the knowledge from other tasks and schemas.

3 IE via Textual Entailment

We first describe how to recast each of the IE tasks (NER, RE, EE, EAE) as TE independently, and leave the workflow between the tasks for the next section. At a high level, the zero-shot TE reformulation consists of three steps: candidate generation, label verbalization and TE inference (Figure 2 illustrates the steps for NER). The first step, candidate generation, identifies text spans (e.g., proper nouns for NER) or span pairs (a pair of entity mentions for relation extraction) in the input sentence as the focus of the prediction. Taking a text span (or span pair) as input, the label verbalization step applies a verbalization template to generate a *hypothesis*, which is a natural language sentence describing the span (or span pair) being an instance of a type of entity, relation, event, or event argument. The verbalization generates hypothesis for each of the target types. Finally, the TE inference step takes the original sentence (the *premise*) and each *hypothesis* as input, and uses a pre-trained TE model to predict if the *premise* entails, contradicts, or is neutral to the *hypothesis*. The type with the verbalization having the highest entailment probability is selected. We next describe each step in detail.

3.1 Candidate Generation

We describe the candidate generation for each of the task below.

Named Entity Recognition (NER): Candidates are extracted using specific patterns of PoS tags as returned by Stanza (Qi et al., 2020). For instance, for the simple example in Figure 1 it suffices to select proper nouns (shown in Figure 2), which are easily extended with other PoS patterns if needed. The toolkit also allows the usage of a constituency parser (Kitaev and Klein, 2018).

Relation Extraction (RE): Each relation requires a pair of entities that satisfy specific type constraints, e.g. the EMPLOYEEOF relation requires a PERSON and an ORGANIZATION. A NER module is used to extract all candidate entities that follow the required entity types according to the target schema. The toolkit uses the TE based NER module, although it also allows usage of a supervised NER system (Qi et al., 2020).

Event (Trigger) Extraction (EE): The main goal of this task is to detect whether the input sentence contains a mention of any of the target event types in the schema, e.g. LIFE.DIE. This task can be formulated as a multi-label text classification task, and in this case the full sentence is the candidate. Alternatively, the textual span that most likely expresses the event (the so-called trigger) can be extracted. In this case, the candidates are generated using specific PoS tags, e.g. verbs like *died* (cf. Figure 1). Our toolkit allows both options.

Event Argument Extraction: Given a sentence containing an event type (as detected by EE above), the goal is to extract entity mentions that are fillers of the target arguments in the schema. For example, the schema in Figure 1 involves three target arguments. Each of the arguments requires specific entity types, e.g. PERSON for the VICTIM argument. The candidates of the required types are extracted using the same NER module as for RE.

3.2 Label Verbalization

For each of the IE tasks, the label verbalization process takes a sentence, a set of candidates and the set of target types (e.g. NER types), and generates a natural language text (the hypothesis) describing the existence of the type in the sentence (the premise) using verbalization templates. Each candidate is a span (or pair of spans) that can belong to a specific type (e.g. being a PERSON in NER). Therefore, the textual verbalization is generated to express each potential type for the span or the pair of spans. For the NER and event extraction tasks, each verbalization expresses one potential entity (or event type) for the target candidate. For the relation and event argument extraction tasks, the verbalization template combines the information from the text spans of the candidate pair and produces a text that expresses a relation (or event argument role). The analyst just needs to write the verbalization templates for each target type, and they are applied to the candidates to generate the hypothesis, as shown in the second step in Figure 2 for NER.

Figure 1 shows sample TE verbalization templates for entity, relation, event, and event argument types corresponding to the 4 IE tasks, as well as sample example as output. The templates for **NER** and **event extraction** (leftmost part of the figure) are applied over a single candidate as extracted in the previous step (the candidate entity or event trigger, respectively). Note that for event extraction it is also possible to produce hypothesis using templates with no slots, e.g. "A person died" for LIFE.DIE. In the case of **relation extraction**, the verbalization templates contain two slots for the two entity spans potentially holding the relation. Finally, templates for **event argument extraction** can be more varied. The figure shows two examples: a template using a single slot for the candidate filler, and a template which, in addition to the filler slot, uses the trigger ("died" in this case, for PLACE).

3.3 Inference

Given a premise (the original sentence) and a hypothesis (an verbalization generated by label verbalization templates), we use a pre-trained TE model to decide whether the hypothesis is entailed by, contradicted with, or is neutral to the premise. In principle, any model trained on an entailment dataset can be used. The inference is mainly determined by three key factors: the TE probabilities for the verbalizations of all templates for all labels, the type-specific input span constraints, and a threshold that decides if the probability is high enough to consider the candidate a positive instance. The type-specific input span constraints are enforced to make sure we don't have candidates that violates the constraints. We return the class label of the hypothesis with highest entailment probability. If none of the hypothesis is higher than the threshold, we return the negative class, that is the class that represents that there is not a valid entity, relation, event, or event argument role type for the input candidate. The threshold for minimal entailment probability is set by default to 0.5.

4 ZS4IE toolkit

ZS4IE comprises a pipeline and a user interface.

4.1 The ZS4IE Pipeline

As described in Section 3.1 and illustrated in Figure 3, there are inter-task dependencies between the four IE tasks (e.g., relation extraction requires that entity mentions have already been tagged in the input sentence). Some task also require external NLP tools for generating candidates. To address these issues and to allow maximal flexibility for the users, we support the following two workflows.

The End-to-End (E2E) Mode: This mode will run the ZS4IE modules in a pipeline: we allow the users to start from raw text, and perform customization (e.g., develop templates for new types of interest) for all four IE tasks. The user has to follow the inter-task dependencies as illustrated in Figure 3: the user must finish NER customization before moving on to relation extraction or the event argument extraction task, because the later two tasks needs NER to generate their input candidates. Similarly, the user must finish customization for the event trigger classification task, before working on the event argument extraction task.

The end-to-end pipeline also runs a customizable pre-processing step including a POS tagger and a constituency parser, before any of the later modules.

The Task Mode: In this mode, the user can choose to work on each of the four IE tasks independently. In order to address the inter-dependencies, the user can choose to run an independent NER module instead, as part of the pre-processing step. The user interface allows the user to tag any spans for entity or event trigger types, before running customization for the more complex tasks such as relation extraction or event argument extraction. This option allows to explore additional entity and event trigger types before actually implementing them

4.2 User Interface (UI)

Figure 4 shows the User Interface. It allows the user to add new types of entities, relations, events and event argument roles, and then develop templates (along with input type constraints for each type). Figure 5 shows the NER extraction results on an user-input sentence. It also displays the likelihood scores produced by the TE model of those

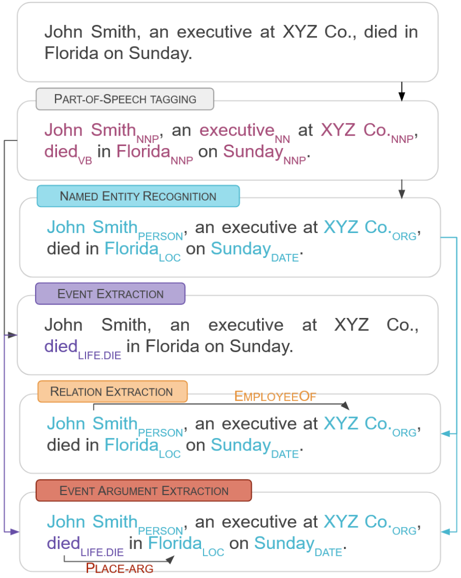

John Smith, an executive at XYZ Co., died in Florida on Sunday.
PART-OF-SPEECH TAGGING
John Smith NNP, an executive NN at XYZ Co. NNP, died VB in Florida NNP on Sunday NNP.
NAMED ENTITY RECOGNITION
John Smith PERSON, an executive at XYZ Co. ORG, died in Florida LOC on Sunday DATE.
EVENT EXTRACTION
John Smith, an executive at XYZ Co., died LIFE.DIE in Florida on Sunday.
RELATION EXTRACTION
EMPLOYEEOF
John Smith PERSON, an executive at XYZ Co. ORG, died in Florida LOC on Sunday DATE.
EVENT ARGUMENT EXTRACTION
John Smith PERSON, an executive at XYZ Co. ORG, died LIFE.DIE in Florida LOC on Sunday DATE.
PLACE-ARG

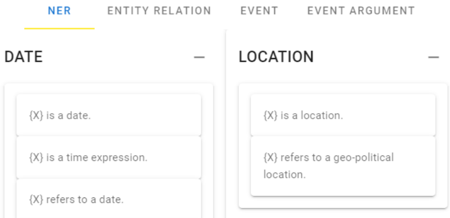
Figure 3: An illustration of the dependencies between the four IE tasks.

templates that are above the threshold, to allow the user to validate templates.

To show why it extracts each entity, it displays a ranked list of likely entity types, the template that led to that type, along with the entailment probability produced by the pre-trained TE model. The user can click on "+" and "-" sign next to each extraction to label its correctness. Our system will track the total number of extractions and and accuracy for each task, each type and each template, to allow the user to quickly validate the effectiveness of the templates and to spot any low-precision template.

Supplying Input Text: The user can supply a text snippet, one at a time, to test writing templates. As described in Section 4.1, when using the task mode, the user can label spans in the input text for the more complex relation extraction and event argument extraction tasks, so that the text already has the right entity or event trigger spans and types to begin with.

Develop Templates for New Types: The user can add new types of entities, relations, events, and event argument role. For each type, the user can create templates along with the input span type constraints, and then run inference interactively on the input text, to see whether these templates



DATE

{X} is a date.

{X} is a time expression.

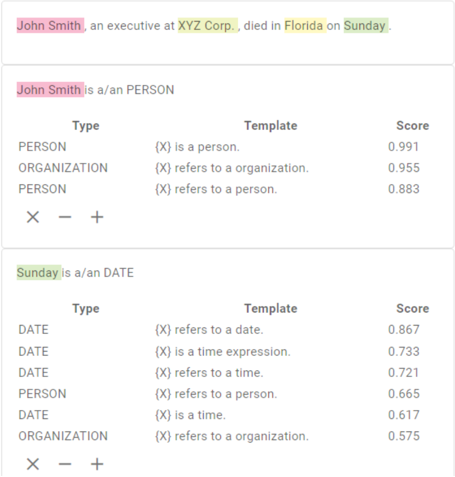
{X} refers to a date.

LOCATION

{X} is a location.

{X} refers to a geo-political location.

Figure 4: The UI for curating templates for types of interests for NER, relation extraction, event extraction and event argument extraction tasks. The NER tab is partially shown with two types.


John Smith, an executive at XYZ Corp., died in Florida on Sunday.

John Smith is a/an PERSON

Type	Template	Score
PERSON	{X} is a person.	0.991
ORGANIZATION	{X} refers to a organization.	0.955
PERSON	{X} refers to a person.	0.883

Sunday is a/an DATE

Type	Template	Score
DATE	{X} refers to a date.	0.867
DATE	{X} is a time expression.	0.733
DATE	{X} refers to a time.	0.721
PERSON	{X} refers to a person.	0.665
DATE	{X} is a time.	0.617
ORGANIZATION	{X} refers to a organization.	0.575

Figure 5: The UI for displaying NER extraction results on an user-input sentence. We show the extractions and the likelihood scores of the templates above the threshold (e.g. $\mathcal{T} = 0.5$).

can be used for extract the instances. The user can label the correctness of the extracted instances, resulting a small development dataset (the *dev* set) to help measuring the precision and relative recall for each template, and to tune the threshold for the TE inference.

Display Metrics: The UI displays the accuracy and yield for each template and each type in real-time, to allow the user to monitor the progress and make adjustments on the fly.

More screenshots and details of our UI are describe in Appendix A.

5 Experiments

We evaluated our system using publicly available datasets. We use CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) for NER evaluation, TACRED (Zhang et al., 2017) for RE, and ACE for EE and EAE (Walker et al., 2006). We evaluate each task independently (not as a pipeline) to make as comparable as possible to existing zero-shot systems. In order to apply our toolkit we made some adaptations as follows: We consider only proper nouns as candidates for NER, and we ignore the MISC label because it is not properly defined in the task [1]. We evaluate EE as event classification, where the task is to output the events mentioned in the sentence without extracting the trigger words, as we found that deciding which is the trigger word is in many cases an arbitrary decision [2]. In the case of RE we used the templates from (Sainz et al., 2021), which are publicly available. We will release the templates used on the experiments as additional material along with the paper. The analysts spent between 5-15 minutes per type, depending on the task, with NER and EE being the fastest.

Table 1 shows the zero-shot results for NER, RE, EE, and EAE tasks. We report the results of three entailment models: RoBERTa (Liu et al., 2019) trained on MNLI, RoBERTa* trained on MNLI, SNLI, FEVER and ANLI; and DeBERTa (He et al., 2021) trained on MNLI. The main results (top three rows) use the default threshold ($\mathcal{T} = 0.5$), we selected the $\mathcal{T}$ blindly, without checking any development result.

The results show strong zero-shot performance. Note that there is no best entailment model, suggesting that there still exists margin for improvement. However, we see that RoBERTa* performs relatively well in all scenarios except EE (see Section 6 for further discussion).

The table also shows in the middle three rows the results where we optimize the threshold on development. The results improve in most of the cases, and allow comparison to other zero-shot systems which sometimes optimize a threshold in development data.

Furthermore, we compare our system with zero-shot task specific approaches from other authors when available. For RE, Wang et al. (2021a) propose a text-to-triple translation method that given a text and a set of entities returns the existing relations. For EE, Lyu et al. (2021) propose, similar

[1] More specifically, we re-labeled the MISC instances to O label.

[2] Note that EAE can be addressed without an explicit mention of the trigger since we used templates that do not require the trigger

	NER			RE			EE			EAE			AVG
Model	Pre	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	F1
RoBERTa	53.3	54.5	53.9	32.8	75.5	45.7	**23.8**	**63.0**	**34.5**	20.5	60.9	30.7	46.7
RoBERTa*	**73.5**	**76.3**	**74.9**	36.8	76.7	49.8	23.5	60.8	33.9	**30.1**	**63.2**	**40.8**	**49.0**
DeBERTa	58.0	50.2	53.8	**40.3**	**77.7**	**53.0**	12.9	60.3	21.2	20.0	31.9	24.6	45.1
RoBERTa (+ $\mathcal{T}$ opt)	49.3	61.8	↑ 54.9	56.1	55.8	↑ 55.9	32.0	52.9	↑ 39.9	25.8	40.1	↑ 31.4	↑ 50.9
RoBERTa* (+ $\mathcal{T}$ opt)	71.9	77.8	↓ 74.8	54.2	59.5	↑ 56.8	25.1	58.6	↑ 35.1	31.1	58.3	↓ 40.6	↑ 51.9
DeBERTa (+ $\mathcal{T}$ opt)	56.3	63.1	↑ 59.5	66.3	59.7	↑ 62.8	13.0	55.8	↓ 21.1	28.9	17.5	↓ 21.8	↑ 51.3
Other authors	-	-	-	-	-	49.2	36.2†	69.1†	47.5†	38.2	35.8	37.0	-

Table 1: Results for NER, RE, EE and EAE experiments results. Three top rows for zero-shot systems with default parameters. Middle rows for threshold optimized on development. The best scores among our results obtained with default thresholds are marked in **bold**. The † indicates non-comparable results due to additional SRL preprocessing.

to us, the use of an entailment model, but in their case the input sentence is split in clauses according to the output of a Semantic Role Labelling system. In order to compare their results with ours, we only use the event types, not the trigger information[3]. The results from our system can be seen as an ablation where we do not make use of any SRL preprocessing. For EAE, Liu et al. (2020) perform zero-shot EAE by recasting the task as QA. Some of these approaches also optimize a threshold on development data, although it is not always clear. We show that our toolkit with default threshold obtains excellent results despite being an all-in-one method.

6 Discussion

Towards post-editing on IE. Our internal evaluation suggest that verbalizing-while-defining workflow can have similar impact as post-editing machine translated text, where human translators obtain quality translations with less effort (Toral et al., 2018). The idea of this new framework will bring down the effort required to create larger and higher quality datasets. Current IE system are subject to a predefined schema and are useless to classify new types of entities, relations and events. The use interface of ZS4IE brings to the annotators the opportunity of defining the schema interactively and manually annotating the dataset with the help of the entailment model. In the future we would like to use the manual annotations to fine-tune the TE model, which would further improve the performance, as shown by the excellent few-shot results of Sainz et al. (2021).

Implicit events extraction. During the development of the EE verbalizations we found out that the entailment model is prone to predict implicit events that are implied by other events. For example, an event type of JUSTICE:JAIL implies an event of JUSTICE:CONVICT where as the same time it implies event type of JUSTICE:TRIAL-HEARING. As the entailment models are not specifically trained for a particular IE task (e.g. EE) they are not limited to the extraction of **explicit** mentions of types (e.g. event types) annotated in the dataset. We think that this phenomenon might have penalized the RoBERTa* model on the EE task, as ACE dataset only contains annotations of explicit events. On the contrary, rather than a limitation of our approach, we believe that this is a positive feature that can be exploited by the users.

7 Conclusions

The ZS4IE toolkit allows a novice user to model complex IE schemas, curating simple yet effective templates for a target schema with new types of entities, relations, events, and event arguments. Empirical validation showed that reformulating the IE tasks as an entailment problem is easy and effective, as spending only 5-15 minutes per type allows to achieve very strong zero-shot performance. ZS4IE brings to the users the opportunity of defining the desired schema on the fly. In addition it allows to annotate examples, similar to post editing MT output. Rather than being a finalized toolkit, we envision several exciting directions, such as including further NLP tasks, allowing the user to select custom pre-processing steps for candidate generation and allowing the user to interactively improve the system annotating examples that are used to fine-tune the TE model.

More generally, we would like to extend the inference capability of our models, perhaps acquired from other tasks or schemas (Sainz et al., 2022),

[3] Output kindly provided by the authors.

in a research avenue where entailment and task performance improve in tandem.

Acknowledgements

Oscar is funded by a PhD grant from the Basque Government (PRE_2020_1_0246). This work is based upon work partially supported via the IARPA BETTER Program contract No. 2019-19051600006 (ODNI, IARPA), and by the Basque Government (IXA excellence research group IT1343-19).

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Qing Lyu, Hongming Zhang, Elior Sulem, and Dan Roth. 2021. Zero-shot event extraction via transfer learning: Challenges and insights. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 322–332, Online. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Towards a unified natural language inference framework to evaluate sentence representations. *CoRR*, abs/1804.08207.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Agirre Eneko. 2022. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In *Findings of the Association for Computational Linguistics: NAACL-HLT 2022*, Online and Seattle, Washington. Association for Computational Linguistics.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oscar Sainz and German Rigau. 2021. Ask2Transformers: Zero-shot domain labelling with pretrained language models. In *Proceedings of the 11th Global Wordnet Conference*, pages 44–52, University of South Africa (UNISA). Global Wordnet Association.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Antonio Toral, Martijn Wieling, and Andy Way. 2018. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, 5:9.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2021a. Zero-shot information extraction as a unified text-to-triple translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1225–1238, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021b. Entailment as few-shot learner.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2856–2878, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A User Interface

We present more details on our user interface (UI) in this section. Our system supports all 4 IE tasks into a single integrated interface.

Template development. Figure 6a shows the main template development UI, in which each tab on the top represents one of the entity, relation, event, and event argument tasks. The user switch between tasks by simply clicking on a different tab (the tabs for the other 3 tasks are shown in Figure 6b, 6c, and 6d, respectively).

Take the NER task as an example (Figure 6a), it shows an overview of all entity types along with the templates defined for each type (e.g., “X is a person” for the type PERSON, in which “X” is a placeholder that can be replaced with a noun phrase “New York City”). If the user clicks on the edit button (the pen-shaped button), the pop-up window for adding a new entity type (the right-hand side figure in Figure 6a) shows up. The user can add a template by clicking on "+" sign, and then input the template to the left (the user can repeat this several times to add more templates). The user can remove a template by clicking on "-". The user can also click on the big "+" card to the left to add a new entity type.

Template development for the relation extraction task is similar to NER, except for two differences: first, as shown in Figure 6b (right), we can further add a set of "allowed type" pairs, that are the set of entity pairs each relation is defined over. For example, the “per:date_of_death” relation is only valid between a pair of PERSON and DATE mentions. Our UI allows the user to specify the “LeftEntityType” (left entity type) and the “RightEntityType” (right entity type) for each relation type under "allowed type". These type constraints are show on the top box for each relation card on the left figure in Figure 6b (e.g., "PERSON->DATE" under “per:date_of_death”). Second, a relation involves a pair of entity mentions. Therefore, each pattern has two placeholders, “X” and “Y”, which can be replaced with two entity candidates that are likely to participate in the relationship.

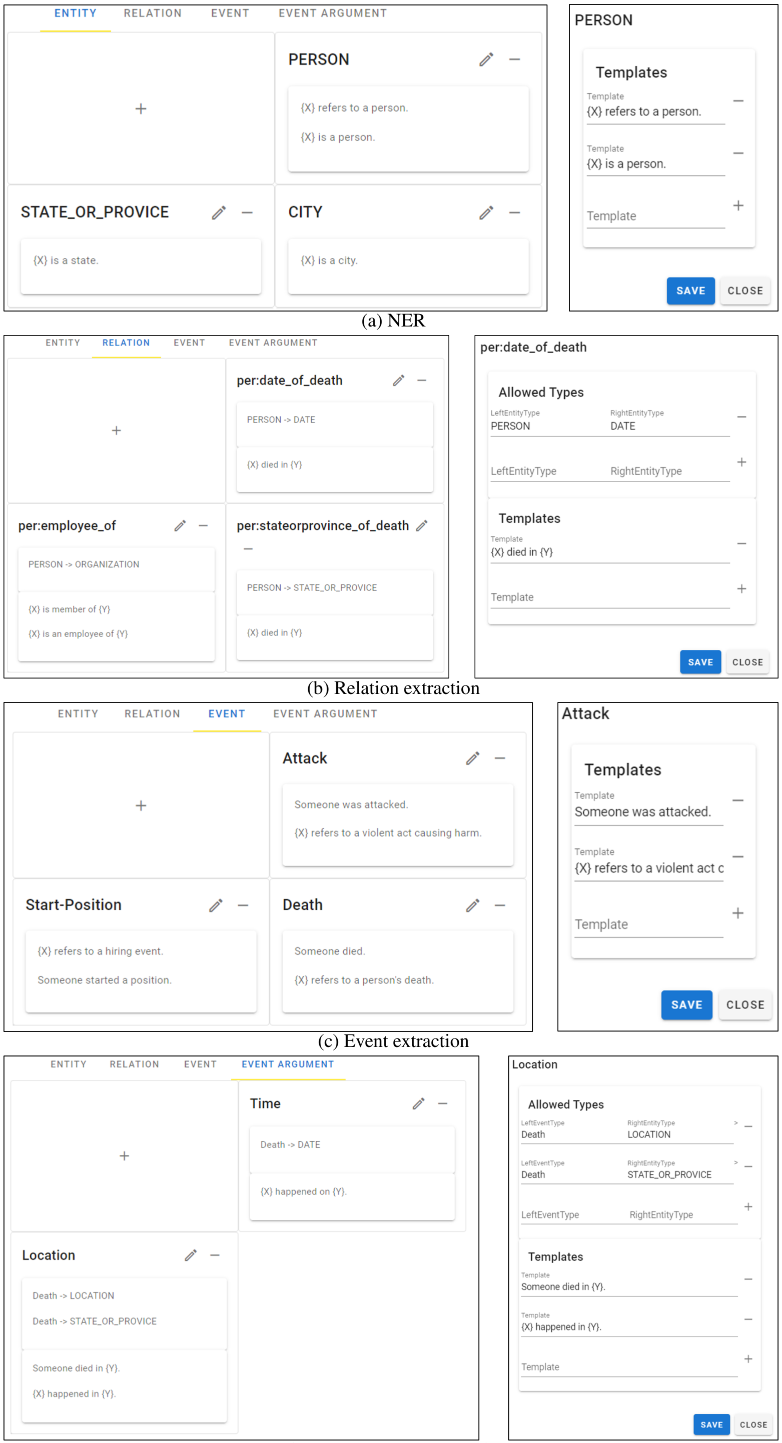
Template development for the event extraction task (Figure 6c) is also similar to NER, except that the template may not contain any trigger. For example, “Someone died” is a template for the “Death” event (Figure 6c). This template would allow the TE approach to classify whether an extent (e.g., a sentence) expresses a type of event.

Template development for the event argument extraction task (Figure 6d) is similar to relation extraction, except that the template can include either two placeholders “X” and “Y” in which “X” is an event trigger and “Y” is an event argument candidate filler (an entity), or only one placeholder “Y” which is the event argument candidate filler. The later would require the template to implicitly describes the event type as well (for example, “Someone died in Y” for the LOCATION event argument role in Figure 6d).

Template validation. We developed an interactive workflow to allow the user to quickly develop templates and validate their effectiveness in our TE-based framework. To support this workflow, our UI allows the user to run inference over any free text supplied by the user herself/himself. For simplicity, we omit the UI where we allow the user type in free text. We show the UI that displays the extraction output on the free text, that also allows the user to label the correcness of the extractions. Based on those labeled examples, the UI also automatically calculate a few metrics to help the user to find the effectiveness of the templates curated so far.

Figure 7 shows the UI for displaying NER extraction outputs (left) and automatically calculated metrics (right). Taken the user-supplied sentence “John Smith, an executive at XYZ Corp., died in Florida on Sunday” as input, the UI on the left-hand side shows the extracted named entities. It shows extractions such as “John Smith is a/an PERSON”, “Sunday is a/an DATE”, and so on. To provide rationale for each extraction, it displays a rank list of possible entity types, the template led to that type, along with the entailment probability produced by the pre-trained TE model. The user can click on "+" and "-" sign next to each extraction to label its correctness. In Figure 7, all extractions are green (labeled by the user as correct) except that “Florida is a/an CITY” is in red (labeled as incorrect by the user). Based on these user-labeled extractions, the system calculated a number of metrics to facilitate template validation: the total number of extracted named entities (shown under “total”), the number of correct and incorrect extractions under “correct” and “incorrect”, respectively (the accuracy number is also shown in the parenthesis next to “correct”) for the overall task, each type, and each pattern. The right-hand side UI in Figure 7 displays these metrics, and allows the user to sort patterns/types by each of the metric. The user can quickly identify some templates are low-precision (e.g., “X is a location” for the entity type CITY), and can revise them to improve precision.

Figure 8a, 8b, and 8c shows the UI for displaying extraction results for the relation extraction, event extraction, and event argument extraction, respectively. Similar to the NER task. Similarly, our system also includes metric board (the metrics above) for the other 3 IE tasks. To view the metric boards for these tasks, please refer to our demonstration video.



PERSON

{X} refers to a person.

{X} is a person.

STATE_OR_PROVICE

{X} is a state.

CITY

{X} is a city.

PERSON

Templates

Template: {X} refers to a person.

Template: {X} is a person.

Template

(a) NER

per:date_of_death

PERSON -> DATE

{X} died in {Y}

per:employee_of

PERSON -> ORGANIZATION

{X} is member of {Y}

{X} is an employee of {Y}

per:stateorprovince_of_death

PERSON -> STATE_OR_PROVICE

{X} died in {Y}

per:date_of_death

Allowed Types

LeftEntityType: PERSON — RightEntityType: DATE

LeftEntityType — RightEntityType

Templates

Template: {X} died in {Y}

Template

(b) Relation extraction

Attack

Someone was attacked.

{X} refers to a violent act causing harm.

Start-Position

{X} refers to a hiring event.

Someone started a position.

Death

Someone died.

{X} refers to a person's death.

Attack

Templates

Template: Someone was attacked.

Template: {X} refers to a violent act c

Template

(c) Event extraction

Time

Death -> DATE

{X} happened on {Y}.

Location

Death -> LOCATION

Death -> STATE_OR_PROVICE

Someone died in {Y}.

{X} happened in {Y}.

Location

Allowed Types

LeftEventType: Death — RightEntityType: LOCATION

LeftEventType: Death — RightEntityType: STATE_OR_PROVICE

LeftEventType — RightEntityType

Templates

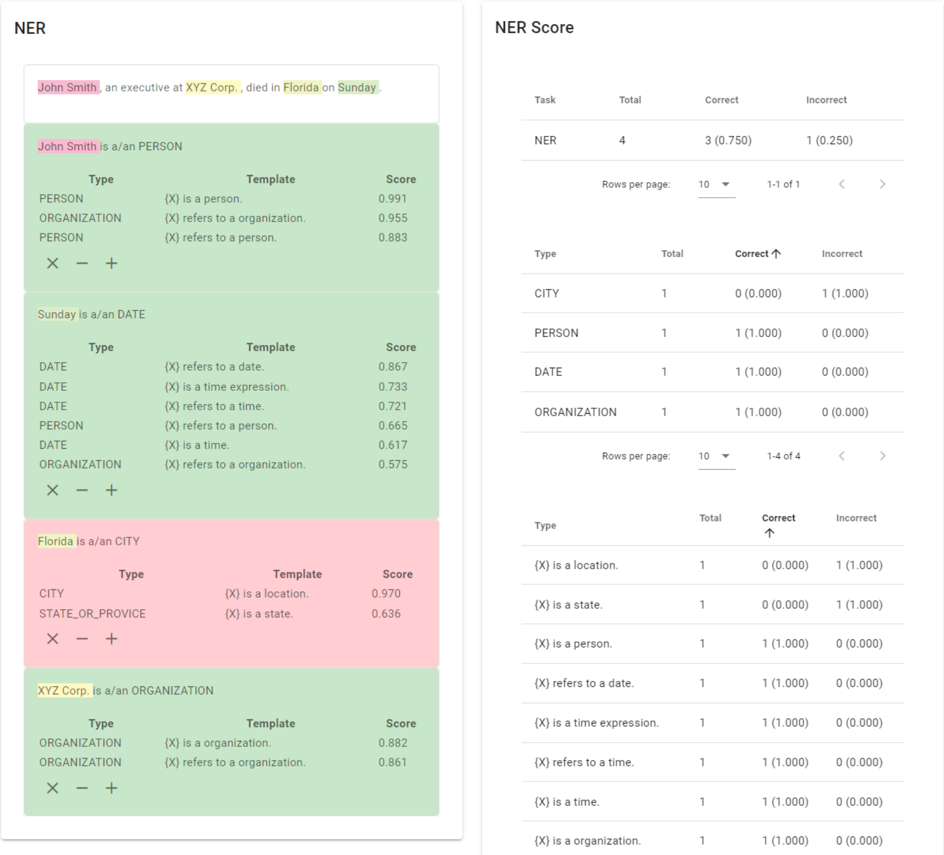
Template: Someone died in {Y}.

Template: {X} happened in {Y}.

Template

(d) Event argument extraction

Figure 6: The UI for developing templates for the 4 IE tasks. For each task, we show the overall UI on the left, and the pop-up window for adding a new entity type PERSON on the right.



NER

John Smith , an executive at XYZ Corp. , died in Florida on Sunday .

John Smith is a/an PERSON

Type	Template	Score
PERSON	{X} is a person.	0.991
ORGANIZATION	{X} refers to a organization.	0.955
PERSON	{X} refers to a person.	0.883

× − +

Sunday is a/an DATE

Type	Template	Score
DATE	{X} refers to a date.	0.867
DATE	{X} is a time expression.	0.733
DATE	{X} refers to a time.	0.721
PERSON	{X} refers to a person.	0.665
DATE	{X} is a time.	0.617
ORGANIZATION	{X} refers to a organization.	0.575

× − +

Florida is a/an CITY

Type	Template	Score
CITY	{X} is a location.	0.970
STATE_OR_PROVICE	{X} is a state.	0.636

× − +

XYZ Corp. is a/an ORGANIZATION

Type	Template	Score
ORGANIZATION	{X} is a organization.	0.882
ORGANIZATION	{X} refers to a organization.	0.861

× − +

NER Score

Task	Total	Correct	Incorrect
NER	4	3 (0.750)	1 (0.250)

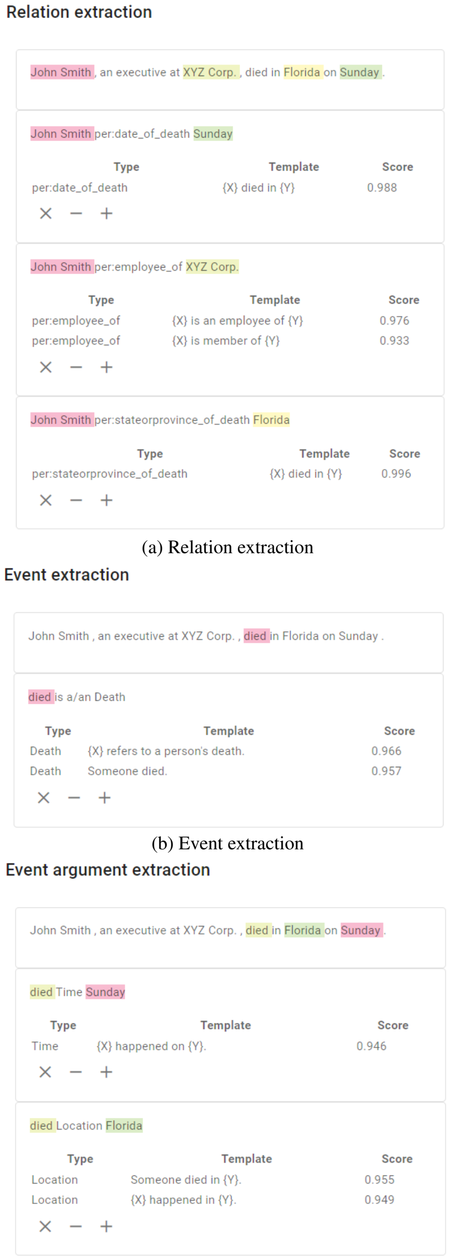
Rows per page: 10 ▾ 1-1 of 1 ‹ ›

Type	Total	Correct ↑	Incorrect
CITY	1	0 (0.000)	1 (1.000)
PERSON	1	1 (1.000)	0 (0.000)
DATE	1	1 (1.000)	0 (0.000)
ORGANIZATION	1	1 (1.000)	0 (0.000)

Rows per page: 10 ▾ 1-4 of 4 ‹ ›

Type	Total	Correct ↑	Incorrect
{X} is a location.	1	0 (0.000)	1 (1.000)
{X} is a state.	1	0 (0.000)	1 (1.000)
{X} is a person.	1	1 (1.000)	0 (0.000)
{X} refers to a date.	1	1 (1.000)	0 (0.000)
{X} is a time expression.	1	1 (1.000)	0 (0.000)
{X} refers to a time.	1	1 (1.000)	0 (0.000)
{X} is a time.	1	1 (1.000)	0 (0.000)
{X} is a organization.	1	1 (1.000)	0 (0.000)

Figure 7: The UI for displaying NER extraction outputs (left) and automatically calculated metrics (right). The left-hand side shows the named entities extracted from an user-input sentence (shown on the top). The user can click on "+" and "-" sign next to each extraction to label its correctness. The right-hand side shows the total number of extracted named entities (total), the number of correct and incorrect extractions (the accuracy number is also shown in the parenthesis next to “correct”) for the overall task, each type, and each pattern. These metrics are calculated based on the set of user labels.



Relation extraction

John Smith , an executive at XYZ Corp. , died in Florida on Sunday .

John Smith per:date_of_death Sunday

Type	Template	Score
per:date_of_death	{X} died in {Y}	0.988

John Smith per:employee_of XYZ Corp.

Type	Template	Score
per:employee_of	{X} is an employee of {Y}	0.976
per:employee_of	{X} is member of {Y}	0.933

John Smith per:stateorprovince_of_death Florida

Type	Template	Score
per:stateorprovince_of_death	{X} died in {Y}	0.996

(a) Relation extraction

Event extraction

John Smith , an executive at XYZ Corp. , died in Florida on Sunday .

died is a/an Death

Type	Template	Score
Death	{X} refers to a person's death.	0.966
Death	Someone died.	0.957

(b) Event extraction

Event argument extraction

John Smith , an executive at XYZ Corp. , died in Florida on Sunday .

died Time Sunday

Type	Template	Score
Time	{X} happened on {Y}.	0.946

died Location Florida

Type	Template	Score
Location	Someone died in {Y}.	0.955
Location	{X} happened in {Y}.	0.949

(c) Event argument extraction

Figure 8: The UI for displaying extractions for relation extraction, event extraction, and event argument extraction, respectively. The user an click on the "+" or "-" sign next to each extraction to label the extraction as correct or incorrect.

GoLLIE : Annotation Guidelines improve Zero-Shot Information-Extraction

Oscar Sainz[*], **Iker García-Ferrero**[*]
Rodrigo Agerri, **Oier Lopez de Lacalle**, **German Rigau**, **Eneko Agirre**
HiTZ Basque Center for Language Technology - Ixa NLP Group
University of the Basque Country (UPV/EHU)
{oscar.sainz, iker.garciaf}@ehu.eus

Abstract

Large Language Models (LLMs) combined with instruction tuning have made significant progress when generalizing to unseen tasks. However, they have been less successful in Information Extraction (IE), lagging behind task-specific models. Typically, IE tasks are characterized by complex annotation guidelines which describe the task and give examples to humans. Previous attempts to leverage such information have failed, even with the largest models, as they are not able to follow the guidelines out-of-the-box. In this paper we propose GoLLIE (**G**uideline-f**o**llowing **L**arge **L**anguage Model for **IE**), a model able to improve zero-shot results on unseen IE tasks by virtue of being fine-tuned to comply with annotation guidelines. Comprehensive evaluation empirically demonstrates that GoLLIE is able to generalize to and follow unseen guidelines, outperforming previous attempts at zero-shot information extraction. The ablation study shows that detailed guidelines is key for good results. Code, data and models are publicly available: https://github.com/hitz-zentroa/GoLLIE.

1 Introduction

The task of Information Extraction (IE) is highly challenging. This challenge is evident in the detailed guidelines, which feature granular definitions and numerous exceptions, that human annotators must follow to perform the task. The performance of current SoTA models heavily depends on the quantity of human-annotated data, as the model learns the guidelines from these examples. However, this performance significantly decreases when tested in new annotation schema (Liu et al., 2021a). The common practice in IE to achieve good results is to manually annotate in each new domain and schema from scratch, as almost no transfer exists across application domains. Unfortunately, this is unfeasible, both, in terms of financial cost and human effort.

Recent advancements in Large Language Models (LLM) (Min et al., 2023) have enabled the development of models capable of generalizing to unseen tasks. Thus, current zero-shot IE systems

[*] Equal contribution

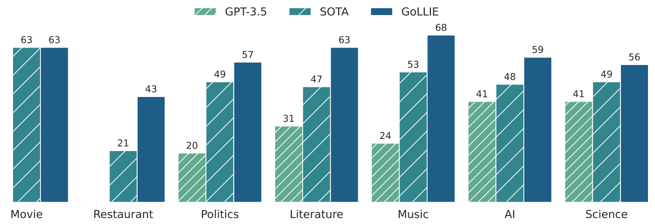


Figure 1: Out of domain zero-shot NER results. GPT results are not available for all domains.

leverage the knowledge encoded in LLMs to annotate new examples (Sainz et al., 2022a; Wang et al., 2023a). As a by product of the pre-training process, models possess now a strong representation of what a person or an organization is. Therefore, they can be prompted to extract mentions to those categories from a text. However, this has a clear limitation: not every annotation schema[*] defines "person" (or any other label) in the same way. For example, ACE05 (Walker et al., 2006) annotates pronouns as persons, while, CoNLL03 (Tjong Kim Sang & De Meulder, 2003) does not. IE tasks require more information than just label names, they require annotation guidelines.

Current LLMs have been trained to follow instructions, but they fail to follow annotation guidelines out-of-the-box. For instance, Figure 1 shows results on domain specific zero-shot Named Entity Recognition. The results of gpt-3.5-turbo when prompted with guidelines (Zhang et al., 2023a) are low, around 20 F1 score on Music or Politics domains. Building a system that enables high-performance zero-shot information extraction, reducing the dependence on costly human annotations, remains an open challenge.

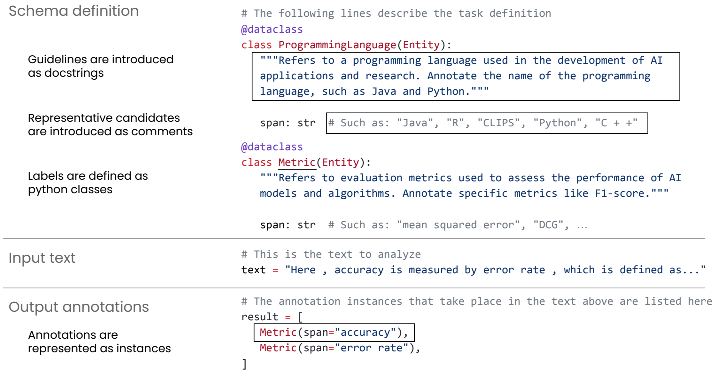
In this work, we present GoLLIE (**G**uideline-f**o**llowing **L**arge **L**anguage Model for **IE**), a LLM fine-tuned to learn how to attend to the guidelines on a small set of well known IE tasks. Comprehensive zero-shot evaluation empirically demonstrates that GoLLIE outperforms the SoTA (Wang et al., 2023a) in zero-shot information extraction (see Figure 1).

2 Related Work

Large Language Models (LLMs) have made significant advancements toward the development of systems that can generalize to unseen tasks (Min et al., 2023). Radford et al. (2019) trained LLMs using a vast amount of internet data, finding that, pretrained models given natural language task descriptions can perform tasks such as question answering, machine translation or summarizing without explicit supervision. Building on this discovery, instruction tuning, often referred to as multitask fine-tuning, has emerged as the leading method to achieve generalization to unseen tasks. This process involves pre-training a model on a massive amount of unlabeled data and subsequently fine-tuning it on a diverse collection of tasks (Wang et al., 2022; Chung et al., 2022) phrased as text-to-text problems (Raffel et al., 2020). A natural language instruction or prompt is given to the model to identify the task it should solve (Schick & Schütze, 2021; Scao & Rush, 2021). Research has demonstrated that increasing the parameter count of the language model (Brown et al., 2020), coupled with improvements in the size and quality of the instruction tuning dataset, results in enhanced generalization capabilities (Chen et al., 2023; Zhang et al., 2022; Chowdhery et al., 2022; Muennighoff et al., 2023; Touvron et al., 2023a;b). LLMs have displayed impressive zero-shot generalization capabilities in various challenging tasks, including coding Wang & Komatsuzaki (2021); Black et al. (2022); Rozière et al. (2023), common sense reasoning Touvron et al. (2023a), and medical applications Singhal et al. (2023), among others.

In the field of Information Extraction (IE), recent shared tasks (Fetahu et al., 2023) have shown that encoder-only language models such as XLM-RoBERTa (Conneau et al., 2020) and mDEBERTA (He et al., 2023) remain the most effective models. Attempts to utilize LLMs and natural language instructions for IE have been less successful (Tan et al., 2023; Zhou et al., 2023; Zhang et al., 2023a), as their performance lags behind that of encoder-only models. Before the billion parameters LLMs, indirectly supervised methods improve zero-shot IE by utilizing the knowledge learned from tasks like Textual Entailment (Sainz et al., 2021; 2022a;b) and Question Answering (Levy et al., 2017). Obeidat et al. (2019) propose an entity typing method that encode label descriptions from Wikipedia as embeddings using an LSTM, which are then used to score the inputs. Methods that leveraged external knowledge were also successful on fine-grained zero-shot NER (Chen et al., 2021). Lu et al. (2022a) introduced a unified text-to-structure generation that can model different IE tasks universally. Lou et al. (2023) proposed converting IE tasks to a semantic matching problem, allowing their method to generalize to new domains and label ontologies not seen during training. Wang et al. (2023a) framed IE tasks as natural language descriptive instructions and trained an LLM across a diverse range of IE tasks. In evaluations on tasks with unseen label ontologies, their model outperformed other instruction-tuning methods.

[*] We define schema as the set of labels and their definitions.


Schema definition
Guidelines are introduced as docstrings
Representative candidates are introduced as comments
Labels are defined as python classes
Input text
Output annotations
Annotations are represented as instances

```python
# The following lines describe the task definition
@dataclass
class ProgrammingLanguage(Entity):
    """Refers to a programming language used in the development of AI
    applications and research. Annotate the name of the programming
    language, such as Java and Python."""

    span: str  # Such as: "Java", "R", "CLIPS", "Python", "C + +"

@dataclass
class Metric(Entity):
    """Refers to evaluation metrics used to assess the performance of AI
    models and algorithms. Annotate specific metrics like F1-score."""

    span: str  # Such as: "mean squared error", "DCG", ...

# This is the text to analyze
text = "Here , accuracy is measured by error rate , which is defined as..."

# The annotation instances that take place in the text above are listed here
result = [
    Metric(span="accuracy"),
    Metric(span="error rate"),
]
```

Figure 2: Example of the input and output of the model.

Most instruction tuning attempts for IE share a limitation: they only consider label names in the prompts (e.g., *"List all the Persons"*). This poses two major challenges. Firstly, not all datasets share the same definition for labels like *Person* (some exclude fictional characters or pronouns). Secondly, a label name alone doesn't sufficiently describe complex or less common labels. While there have been attempts to prompt LLMs using guidelines (Zhang et al., 2023a), strong prior knowledge of LLMs regarding task labels (Blevins et al., 2023) deter the model from adhering to those guidelines.

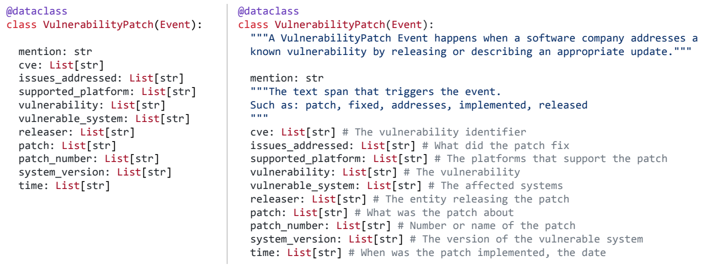
3 APPROACH

Different from previous approaches, GoLLIE forces the model to attend to the details in the guidelines, performing robustly on schemas not seen during training. On this section we deep dive into the details of our approach, describing how the input and output was represented and the regularization techniques used to force the model to attend to the guidelines.

3.1 INPUT-OUTPUT REPRESENTATION

We have adopted a Python code-based representation (Wang et al., 2023b; Li et al., 2023) for both the input and output of the model. This approach not only offers a clear and human-readable structure but also addresses several challenges typically associated with natural language instructions. It enables the representation of any information extraction task under a unified format. The inputs can be automatically standardized using Python code formatters such as Black. The output is well-structured and parsing it is trivial. Furthermore, most current LLMs incorporate code in their pretraining datasets, indicating that these models are already familiar with this representation.

Figure 2 shows the three main parts of the format: schema definition, input text and output annotations. **Schema definition** forms the initial segment of the input. This section contains the information about the labels which are represented as Python classes; guidelines, articulated as docstrings; and representative annotation candidates presented in the form of code comments. The number of class definitions corresponds to the number of labels in the dataset. Classes are flexible and vary for each task. For example, classes for a NER dataset merely require an attribute to specify the text span that correspond to the class. On the other side, more complex tasks such as Event Argument Extraction (EAE) or Slot Filling (SF) demand more class attributes to categorize the task, such as a list of participants in an event (refer to examples in Appendix A). **Input text** is the second part of the input. The input text is represented as an string variable in Python. **Output annotations** is the part generated by the model. The model starts generating after `result =`. The annotations are represented as a list of instances of the classes defined on the schema definition part. Parsing the output is straightforward; executing the generated code in Python yields a list containing the result. This ease of parsing the output stands as a significant advantage of our model. A further detailed analysis of the efficiency of this approach is available in Appendix E.



```
@dataclass
class VulnerabilityPatch(Event):

  mention: str
  cve: List[str]
  issues_addressed: List[str]
  supported_platform: List[str]
  vulnerability: List[str]
  vulnerable_system: List[str]
  releaser: List[str]
  patch: List[str]
  patch_number: List[str]
  system_version: List[str]
  time: List[str]
```

```
@dataclass
class VulnerabilityPatch(Event):
  """A VulnerabilityPatch Event happens when a software company addresses a
  known vulnerability by releasing or describing an appropriate update."""

  mention: str
  """The text span that triggers the event.
  Such as: patch, fixed, addresses, implemented, released
  """
  cve: List[str] # The vulnerability identifier
  issues_addressed: List[str] # What did the patch fix
  supported_platform: List[str] # The platforms that support the patch
  vulnerability: List[str] # The vulnerability
  vulnerable_system: List[str] # The affected systems
  releaser: List[str] # The entity releasing the patch
  patch: List[str] # What was the patch about
  patch_number: List[str] # Number or name of the patch
  system_version: List[str] # The version of the vulnerable system
  time: List[str] # When was the patch implemented, the date
```

Figure 3: Example of the input representation. (left) An example of an event definition w/o guidelines information. (right) The same example but with guideline information as Python comments.

3.2 Guidelines enhanced representation

The main contribution of this work is the use of the guidelines as part of the inference process to improve the zero-shot generalization. An example of a class definition with and without guidelines is shown in the Figure 3. Different datasets usually define guidelines on many different ways: some provides a complex definition of a label with several exceptions and special treatments and others just give a few representative candidates of the fillers of the label. To normalize the input format, we included the label definitions as class docstrings and the candidates as a comment for the principal argument (which is usually *mention* or *span*). Complex tasks such as EAE or SF requires additional definitions for the arguments or slots, to that end, we included small definitions as comments on each class argument. In this paper, we will refer to the model without guidelines as Baseline and the model with guidelines as GoLLIE.

3.3 Training regularization

We want to ensure that the model follows the guidelines and does not just learn to identify specific datasets and perform correctly on them. To do this, we introduce various kinds of noise during training. This stops the model from recognizing particular datasets, recalling specific labels, or attending only to the label names rather than learning to follow the actual description for each label in the guidelines.

We applied the following regularizations. **Class order shuffling**, for each example, the order of the input classes is randomly shuffled. This makes it more difficult for the model to memorize entire task definitions. **Class dropout**, we delete some of the input classes randomly. By eliminating few classes from both the input and output, we force the model to learn to only output instances of classes defined in the input. This not only encourages the model to focus on the schema definition but also minimizes the occurrence of hallucinations during inference. **Guideline paraphrasing**, we generate variations of the label definitions to prevent the model from easily memorizing them. We also think this will make the method more robust to different variations on the definition. **Representative candidate sampling**, similar to what we do with the paraphrases, for each input we sample 5 different candidates from a fixed pool of 10 per class. **Class name masking** involves substituting the label class names (e.g., PERSON) with placeholders, such as LABEL_1. This prevents the model from exploiting the label names during training, and forces it to attend and understand the guidelines.

4 Experimental Setup

4.1 Data

Evaluating zero-shot capabilities requires dividing the data into training and evaluation datasets. However, many benchmarks for Information Extraction are based on the same domain or share part of their schema. To ensure that the zero-shot evaluation is not affected by similar data, we have divided our set of benchmarks based on the domain of the data (a related topic is data contamination,

Table 1: Datasets used on the experiments. The table shows the domain, tasks and whether are use for training, evaluation or both.

Dataset	Domain	NER	RE	EE	EAE	SF	Training	Evaluation
ACE05 (Walker et al., 2006)	News	✓	✓	✓	✓		✓	✓
BC5CDR (Wei et al., 2016)	Biomedical	✓					✓	✓
CoNLL 2003 (Tjong Kim Sang & De Meulder, 2003)	News	✓					✓	✓
DIANN (Fabregat et al., 2018)	Biomedical	✓					✓	✓
NCBIDisease (Islamaj Doğan & Lu, 2012)	Biomedical	✓					✓	✓
Ontonotes 5 (Pradhan et al., 2013)	News	✓					✓	✓
RAMS (Ebner et al., 2020)	News				✓		✓	✓
TACRED (Zhang et al., 2017)	News					✓	✓	✓
WNUT 2017 (Derczynski et al., 2017)	News	✓					✓	✓
BroadTwitter (Derczynski et al., 2016)	Twitter	✓						✓
CASIE (Satyapanich et al., 2020)	Cybercrime			✓	✓			✓
CrossNER (Liu et al., 2021b)	*Many*	✓						✓
E3C (Magnini et al., 2021)	Biomedical	✓						✓
FabNER (Kumar & Starly, 2022)	Science	✓						✓
HarveyNER (Chen et al., 2022)	Twitter	✓						✓
MIT Movie (Liu et al., 2013)	Queries	✓						✓
MIT Restaurants (Liu et al., 2013)	Queries	✓						✓
MultiNERD (Tedeschi & Navigli, 2022)	Wikipedia	✓						✓
WikiEvents(Li et al., 2021)	Wikipedia	✓		✓	✓			✓

that we discuss in Appendix G). For training we kept mostly datasets from **News and Biomedical** domains, for evaluation instead, we used datasets from **diverse domains**. This approach helps to avoid introducing any noise into the evaluation process. Among the evaluation datasets we included CrossNER (Liu et al., 2021b), a dataset that is split into many domains, for simplicity, we will call each domain as a separate dataset: AI, Literature, Music, Politics and Science. Also, we will call MIT Movie and MIT Restaurant as Movie and Restaurant. Table 1 contains the information about the data used in the experiments.

We have trained the model to perform 5 different tasks: Named Entity Recognition (NER), Relation Extraction (RE), Event Extraction (EE), Event Argument Extraction (EAE) and Slot Filling (SF). However, we only evaluated the model on the three main tasks of interest: NER, EE and EAE. The other two tasks are added in the training data in order to add diversity and improve the flexibility of the model.

Few modifications has been done to two datasets in order to improve the quality of the model. First, the training data of Ontonotes 5 was reduced drastically as it was automatically annotated. Second, the TACRED dataset was converted from RE to SF in order to increase the complexity of the task. These modifications make our system not comparable with the state of the art on those tasks. However, our focus of interest is in the zero-shot evaluation and, therefore, the benefits (see Appendix A) are more interesting than adding 2 more comparable points on the supervised setup. In the CASIE dataset, we detected that the annotated event spans are inconsistent. The models typically annotate a sub-string rather than the entire span. Therefore, we evaluate all the models based on the predicted event categories, without considering the exact text span. For arguments, we use partial matching.

We use the guidelines released by the authors of each dataset (More details in Appendix F). When such guidelines are not publicly available, we ask human experts to create them, based on the annotations from the development split. The representative candidates are extracted from the guidelines when available, otherwise, the candidates are sampled from the the train split based on word frequency or manually curated based on the guidelines. Paraphrases are automatically generated using Vicuna 33B v1.3 (Zheng et al., 2023).

4.2 Language Models and Baselines

Backbone LLMs: GoLLIE is a fine-tuned version of Code-LLaMA Rozière et al. (2023). Other backbone LLMs, such as LLaMA (Touvron et al., 2023a), LLaMA-2 Touvron et al. (2023b) or Falcon Penedo et al. (2023) were considered during the development, however, as our approach uses code to represent the input and output, Code-LLaMA model worked better on the preliminary experiments. In order to perform fair comparisons the baseline developed in this paper is based on Code-LLaMA as well. All the development of this paper was done with the 7B parameter version of Code-LLama, but, for a scaling analysis we also trained the 13B and 34B parameter models.

Table 2: Supervised evaluation results. “*” indicates that results are not directly comparable.

Dataset	SoTA	Baseline	🐶	🐶 13B	🐶 34B
ACE05$_{\text{NER}}$	(Wang et al., 2023a) 86.6	$89.1_{\pm 0.2}$	$88.1_{\pm 0.6}$	$89.4_{\pm 0.2}$	$\mathbf{89.6}_{\pm 0.1}$
ACE05$_{\text{RE}}$	(Lu et al., 2022b) 66.1	$63.8_{\pm 0.6}$	$63.6_{\pm 1.8}$	$67.5_{\pm 0.5}$	$\mathbf{70.1}_{\pm 1.5}$
ACE05$_{\text{EE}}$	(Lu et al., 2022b) **73.4**	$71.7_{\pm 0.2}$	$72.2_{\pm 0.8}$	$70.9_{\pm 1.6}$	$71.9_{\pm 1.1}$
ACE05$_{\text{EAE}}$	(Lu et al., 2022b) *54.8	$65.9_{\pm 0.7}$	$66.0_{\pm 0.8}$	$67.8_{\pm 0.9}$	$\mathbf{68.6}_{\pm 1.2}$
BC5CDR	(Zhang et al., 2023b) **91.9**	$87.5_{\pm 0.2}$	$87.5_{\pm 0.2}$	$87.9_{\pm 0.1}$	$88.4_{\pm 0.2}$
CoNLL 2003	(Lu et al., 2022b) 93.0	$92.9_{\pm 0.1}$	$92.8_{\pm 0.3}$	$93.0_{\pm 0.2}$	$\mathbf{93.1}_{\pm 0.1}$
DIANN	(Zabala et al., 2018) 74.8	$80.3_{\pm 0.7}$	$79.4_{\pm 1.1}$	$82.6_{\pm 1.3}$	$\mathbf{84.1}_{\pm 1.1}$
NCBIDisease	(Wang et al., 2023a) **90.2**	$86.2_{\pm 0.1}$	$85.4_{\pm 0.3}$	$86.5_{\pm 0.8}$	$85.8_{\pm 0.2}$
Ontonotes 5	-	$83.4_{\pm 0.2}$	$83.4_{\pm 0.2}$	$84.0_{\pm 0.2}$	$\mathbf{84.6}_{\pm 0.4}$
RAMS	(Li et al., 2021) 48.6	$48.9_{\pm 0.4}$	$48.7_{\pm 0.7}$	$49.6_{\pm 0.1}$	$\mathbf{51.2}_{\pm 0.3}$
TACRED	-	$56.6_{\pm 0.2}$	$57.1_{\pm 0.9}$	$56.7_{\pm 0.5}$	$\mathbf{58.7}_{\pm 0.2}$
WNUT 2017	(Wang et al., 2021) **60.2**	$53.7_{\pm 0.7}$	$52.0_{\pm 0.6}$	$50.5_{\pm 0.9}$	$54.3_{\pm 0.4}$
Average		$73.3_{\pm 0.1}$	$73.0_{\pm 0.3}$	$73.9_{\pm 0.3}$	$\mathbf{75.0}_{\pm 0.3}$

Training setup: To train the models we use QLoRA (Hu et al., 2022; Dettmers et al., 2023). LoRA freezes the pre-trained model weights and injects trainable rank decomposition matrices into linear layers of the Transformer architecture. On a preliminary experiment this setup outperformed fine-tuning the entire model on the zero-shot tasks, while trained much faster (more details in Appendix D.4). We applied the LoRA to all linear transformer block layers as recommended by Dettmers et al. (2023). The models were trained for 3 epochs with an effective batch-size of 32 and a learning-rate of 3e-4 with cosine scheduler. Our training infrastructure was 2 NVIDIA's A100 with 80gb each. More details about the training are given in the Appendix D.

Comparable systems: Our main point of comparison is Instruct-UIE (Wang et al., 2023a) as it is the approach closest to our system, but does not use guidelines. Another system considered for comparison is PromptNER (Zhang et al., 2023a), which proposes to prompt GPT-3.5 and T5 with definitions using Chain-of-Though in order to perform few-shot NER. Different from us, they did not fine-tune the model to attend to the guidelines. For fair comparison, we only considered the zero-shot results reported in the paper. In addition, other SoTA systems are added for comparison when results from Instruct-UIE and PromptNER are not available. Given that our systems is designed for the zero-shot scenario, the supervised experiments are intended to verify that our system does not degrade its performance. We thus we selected, for the supervised scenario, those systems among SoTA that share the most comparable setting with us.

5 Results

5.1 Supervised evaluation

The results on the supervised datasets are shown in Table 2. Comparing GoLLIE with the baseline, they both obtain very similar results, with an absolute difference of 0.3 F1 points on average. This is expected, as the baseline model implicitly learns the guidelines for annotating the datasets based on the data distribution during fine-tuning. In addition, despite the noise introduced to GoLLIE fine-tuning in order to generalize from guidelines, the performance is close to that of the baseline.

Compared to other systems our model achieves similar results in general. Focusing on the two datasets where our model under-performs significantly, WNUT and NCBIDisease, we find that task specific techniques are still needed. For instance, Wang et al. (2021) uses external knowledge to detect emergent and rare entities. In the NCBIDisisease dataset, models pre-trained on Biomedical domain corpora achieve best results (Kocaman & Talby, 2021). (Wang et al., 2023a) leverages Flan-T5, which has great proficiency on Biomedical domain tasks (Singhal et al., 2022). These improvements, however, are complementary to our proposal.

5.2 Zero-Shot evaluation

The results on the zero-shot are shown in Table 3. Overall, comparing to the baseline, **the results are improved significantly when using guidelines** on almost every dataset, with an absolute difference

Table 3: Zero-shot evaluation results. “*” indicates results obtained using the original code.

Dataset	SoTA	Baseline	🐶	🐶 13B	🐶 34B
BroadTwitter	-	39.0$_{\pm0.6}$	49.5$_{\pm0.8}$	**51.4**$_{\pm1.8}$	50.3$_{\pm2.1}$
CASIE$_{\text{EE}}$	-	33.9$_{\pm6.5}$	59.3$_{\pm2.3}$	62.2$_{\pm0.9}$	**65.5**$_{\pm1.8}$
CASIE$_{\text{EAE}}$	-	47.9$_{\pm1.4}$	50.0$_{\pm1.1}$	52.6$_{\pm0.2}$	**55.2**$_{\pm0.5}$
AI	(Wang et al., 2023a) 49.0	32.3$_{\pm0.8}$	59.1$_{\pm1.1}$	56.7$_{\pm3.0}$	**61.6**$_{\pm1.9}$
Literature	(Wang et al., 2023a) 47.2	39.4$_{\pm0.7}$	**62.7**$_{\pm3.2}$	59.7$_{\pm0.3}$	59.1$_{\pm2.6}$
Music	(Wang et al., 2023a) 53.2	56.2$_{\pm1.3}$	67.8$_{\pm0.2}$	65.5$_{\pm3.6}$	**68.4**$_{\pm2.1}$
Politics	(Wang et al., 2023a) 48.2	38.3$_{\pm1.1}$	57.2$_{\pm1.0}$	54.4$_{\pm4.1}$	**60.2**$_{\pm3.0}$
Science	(Wang et al., 2023a) 49.3	37.1$_{\pm1.3}$	55.5$_{\pm1.6}$	56.2$_{\pm1.0}$	**56.3**$_{\pm0.4}$
E3C	-	59.8$_{\pm0.3}$	59.0$_{\pm0.7}$	59.0$_{\pm0.8}$	**60.0**$_{\pm0.4}$
FabNER	-	06.1$_{\pm0.4}$	24.8$_{\pm0.6}$	25.4$_{\pm0.5}$	**26.3**$_{\pm0.4}$
HarveyNER	-	23.2$_{\pm0.4}$	37.3$_{\pm1.8}$	**41.3**$_{\pm0.8}$	38.9$_{\pm0.5}$
Movie	(Wang et al., 2023a) 63.0	43.4$_{\pm1.1}$	**63.0**$_{\pm0.6}$	62.5$_{\pm1.0}$	62.4$_{\pm1.4}$
Restaurants	(Wang et al., 2023a) 21.0	31.3$_{\pm2.2}$	43.4$_{\pm0.8}$	49.8$_{\pm1.4}$	**52.7**$_{\pm1.6}$
MultiNERD	-	55.0$_{\pm1.1}$	76.0$_{\pm0.7}$	**77.5**$_{\pm0.3}$	77.2$_{\pm0.6}$
WikiEvents$_{\text{NER}}$	(Sainz et al., 2022b) *49.1	76.9$_{\pm5.1}$	80.7$_{\pm0.7}$	80.2$_{\pm0.7}$	**81.3**$_{\pm0.5}$
WikiEvents$_{\text{EE}}$	(Sainz et al., 2022b) *10.4	47.5$_{\pm0.4}$	43.0$_{\pm0.6}$	45.7$_{\pm0.8}$	**47.0**$_{\pm1.9}$
WikiEvents$_{\text{EAE}}$	Sainz et al. (2022a) 35.9	51.6$_{\pm0.5}$	51.9$_{\pm0.4}$	**52.5**$_{\pm1.2}$	50.7$_{\pm0.4}$
Average SoTA	42.6	45.4$_{\pm0.5}$	58.4$_{\pm0.5}$	58.3$_{\pm0.7}$	**60.0**$_{\pm1.0}$
Average all	-	42.3$_{\pm0.2}$	55.3$_{\pm0.2}$	56.0$_{\pm0.2}$	**57.2**$_{\pm0.5}$

of 13 F1 points on average. Despite dividing the evaluation benchmarks based on the domain, there is always some overlap between labels of train and evaluation benchmarks. For instance, the datasets E3C and WikiEvents share a large part of their schema with datasets like BC5CDR, ACE05 and RAMS. This phenomena is reflected in the results.

GoLLIE surpass by a large margin the current zeri-shot SoTA methods Instruct-UIE (Wang et al., 2023a) and Entailment based IE (Sainz et al., 2022b). Compared to Instruct-UIE, the main differences are the backbone model, the amount of training data, and, the use or not of the guidelines. Instruct-UIE leverages the 11B FlanT5 which is a T5 fine-tuned on 473 NLP datasets. Respect to the data, Instruct-UIE leverages a total of 34 IE datasets (counting different tasks as datasets) from diverse domains, we only leverage 12 datasets. Contrary to our method they do not use guideline information. Still, our method performs significantly better suggesting that the guidelines have an important effect on the results.

PromptNER (Zhang et al., 2023a) also adds some definition information into the prompt in order to perform zero-shot NER. We compare our approach with them (represented as GPT-3.5) in Figure 1. Although their approach leverages guidelines too, our approach performs significantly better on all datasets, showing that LLMs (even with 175B parameters) struggle to follow guidelines. They solve this by adding examples in the context but still far behind on a comparable setting (T5-XXL).

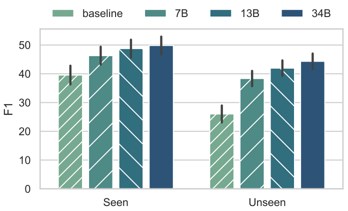


Figure 4: Seen vs unseen label zero-shot performance, results aggregated from all datasets.

Seen vs unseen labels: Not all labels in the zero-shot datasets are unseen; there is an overlap between the labels in the training and zero-shot datasets. Although these labels may have very different annotation guidelines, we also report results on the set of labels to which it has not been exposed during training, to better understand the generalization capabilities of GoLLIE. The list of seen and unseen labels, as well as an extended analysis is available in Appendix B. Figure 4 aggregates the F1 scores across datasets for seen and unseen labels in the zero-shot scenario. All models exhibit slightly lower performance on unseen labels. For the baseline model, the performance drop is more pronounced. In contrast, GoLLIE demonstrates better generalization ability, showing a smaller gap in F1 scores between the seen and unseen labels. Also, the gap is smaller as the parameter count of our model increases.

Model scaling: Recent research has shown that increasing the parameter count of language models leads to improved generalization capabilities Brown et al. (2020). Higher parameter count yields superior average zero-shot performance. However, some datasets and tasks greatly benefit from a larger LLM, while others do not. We believe that some datasets do not see benefits from increasing the LLM size because their performance is hindered by the issues with the guidelines that we discuss in Section 5.3. While, in general, larger models achieve better results in both supervised and zero-shot settings, GoLLIE with a 7B parameter backbone already exhibits strong zero-shot capabilities.

5.3 Ablation study

We have performed an ablation to see the contribution of several components in the zero-shot evaluation. We analyzed the different regularization techniques proposed in Section 3.3. Additionally, we represent the baseline, i.e when removing all components including guidelines, as ”w/o *all*”. Along with the mean zero-shot F1 we also provide the one-sided p-value with respect to 🦙 GoLLIE.

The class order shuffling, guideline paraphrasing and class name masking seems to have no significant contribution to the final result, while class dropout although significant the improvements are small. As further explained in Appendix D, the loss is only computed over the result tokens, inherently limiting the model's potential to overfit to the guidelines. In contrast, the representative annotation items give some stronger signal to the model. We see how definitions and representative candidates from the guidelines are complementary and help to improve each other.

Table 4: Ablation results.

Model	**F1**	**p-value**
🦙 GoLLIE	$55.3_{\pm 0.2}$	-
w/o Shuffling	$55.9_{\pm 0.2}$	$7.2e^{-2}$
w/o Paraphrases	$54.8_{\pm 0.2}$	$1.1e^{-1}$
w/o Masking	$54.6_{\pm 0.6}$	$1.0e^{-1}$
w/o Dropout	$54.0_{\pm 0.2}$	$4.0e^{-3}$
w/o Candidates	$49.9_{\pm 0.2}$	$2.2e^{-10}$
w/o *all* (baseline)	$42.3_{\pm 0.1}$	$5.1e^{-13}$

6 Error analysis

In this section, we aim to better understand the effect of prompting LLMs with guidelines. We focus on specific labels across various datasets, with the results displayed in Table 5. Our analysis covers both successful and unsuccessful cases of entity labeling by GoLLIE. For the latter, we also aim to identify the reasons why the model fails to correctly label these entities. Further analyses on malformed outputs or hallucinations are discussed in Appendix C.

Table 5: This table shows the F1 scores for specific labels from different datasets. The guideline column is a small summary of the actual guideline used to prompt the model.

Dataset	**Label**	**Guideline**	**Baseline**	🦙
MultiNERD	Media	Titles of films, books, songs, albums, fictional characters and languages.	13.6	69.1
CASIE	Vul. Patch	When a software company addresses a vulnerability by releasing an update.	27.7	70.5
Movie	Trailer	Refers to a short promotional video or preview of a movie.	00.0	76.4
AI	Task	Particular research task or problem within a specific AI research field.	02.7	63.9
MultiNERD	Time	Specific and well-defined time intervals, such as eras, historical periods, centuries, years and important days.	01.4	03.5
Movie	Plot	Recurring concept, event, or motif that plays a significant role in the development of a movie.	00.4	05.1
AI	Misc	Named entities that are not included in any other category.	01.1	05.2
Literature	Misc	Named entities that are not included in any other category.	03.7	30.8
Literature	Writer	Individual actively engaged in the creation of literary works.	04.2	65.1
Literature	Person	Person name that is not a writer.	33.5	49.4
Science	Scientist	A person who is studying or has expert knowledge of a natural science field.	02.1	05.8
Science	Person	Person name that is not a scientist.	46.1	45.9
Politics	Polit. Party	Organization that compete in a particular country's elections.	11.2	34.9

The details are in the guidelines: Labels such as Media, VulnerabilityPatch, Trailer, and Task are inherently polysemous, making it challenging to determine the appropriate categorization based solely on the label name. As a result, the baseline struggles to effectively classify items under these labels due to having insufficient information. Conversely, GoLLIE successfully follows the guidelines, underscoring their utility.

When the annotations do not comply with the guidelines: In the case of the Time label of the MultiNERD dataset, we found that our model labels years as Time entities. This is correct according to the annotation guidelines. Surprisingly, years are not labeled as entities in the dataset. In this case, GoLLIE successfully follows the guidelines; unfortunately, the dataset annotations do not.

Ambiguous labels: The Miscellaneous category, used by CoNLL03 and CrossNER datasets, refers to any named entity that is not included in the predefined categories set by the dataset. This definition is highly ambiguous and serves as a catch-all for various elements that do not fit into any of the predefined categories. Similarly, the Plot category of the Movie dataset is used to label a wide range of elements. For example, events in a movie (e.g., murder, horse racing), characters (e.g., vampires, zombies), and the country of origin (e.g., British), among others. This lack of specificity hinders the development of consistent rules or guidelines for tagging such elements (Ratinov & Roth, 2009), which is a problem for humans and machines alike. As a consequence, GoLLIE also fails to label them accurately.

Conflicts Between Fine-Grained and Coarse Entities: The CrossNER dataset introduces two labels for person names within each domain. For example, in the Science domain, the labels Scientist and Person are used. The former, is used to label any person that is not a Scientist. Similarly, the Literature domain includes the labels Writer and Person. The guidelines assist GoLLIE in correctly labeling entities as Writer. However, GoLLIE still categorizes individuals as *Person* even when they are *Scientist*, despite the guidelines. This is not technically incorrect, as every scientist is, by definition, also a person.

Strong Label Preconceptions: In its Political domain set, CrossNER includes the label Political Party. GoLLIE outperforms the baseline, once again demonstrating the utility of providing the model with guidelines. However, we often find that the model categorizes political parties as organizations. As listed in Table 1, most of the pretraining datasets originate from the news domain, where political parties are a common entity. However, none of the fine-tuning datasets include the Political Party entity; they are instead categorized as Organization. Consequently, during inference, the model consistently labels political parties as organizations. We believe this issue can be resolved by expanding the number and diversity of the fine-tuning datasets.

In summary, we anticipate that **GoLLIE will perform well on labels with well-defined and clearly bounded guidelines**. On the other hand, ambiguous labels or very coarse labels pose challenges. To this regard, be believe that GoLLIE would benefit from learning to follow instructions such as *”Label always the most specific class”* or *”Annotate this class in the absence of other specific class”*. We also expect that GoLLIE would benefit from expanding the number and diversity of the pre-training datasets.

7 Conclusions

In this paper we introduce GoLLIE, a LLM specifically fine-tuned to comply with annotation guidelines that were devised for helping humans to annotate the dataset. A comprehensive zero-shot evaluation empirically demonstrate that annotation guidelines are of great value for LLMs, as GoLLIE successfully leverages them. GoLLIE achieves better zero-shot results than previous attempts at zero-shot IE which do not leverage the guidelines, or use models not finetuned for following guidelines.

GoLLIE is a significant progress towards the development of models that can generalize to unseen IE tasks. In the future, we plan to enhance GoLLIE by using a larger and more diverse set of pre-training datasets. We will also improve the model's performance with ambiguous and coarse labels by expanding the set of instructions that the model can follow.

ACKNOWLEDGMENTS

This work has been partially supported the Basque Government (Research group funding IT-1805-22). We are also thankful to several MCIN/AEI/10.13039/501100011033 projects: (i) DeepKnowledge (PID2021-127777OB-C21) and ERDF A way of making Europe; (ii) Disargue (TED2021-130810B-C21) and European Union NextGenerationEU/PRTR; (iii) AWARE (TED2021-131617B-I00) and European Union NextGenerationEU/PRTR. This work has also been partially funded by the LUMINOUS project (HORIZON-CL4-2023-HUMAN-01-21-101135724). Oscar Sainz is supported by a doctoral grant from the Basque Government (PRE_2023_2_0137). Rodrigo Agerri currently holds the RYC-2017-23647 fellowship (MCIN/AEI/10.13039/501100011033 and by ESF Investing in your future).

REFERENCES

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In Angela Fan, Suzana Ilic, Thomas Wolf, and Matthias Gallé (eds.), *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136, virtual+Dublin, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bigscience-1.9. URL https://aclanthology.org/2022.bigscience-1.9.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. Prompting language models for linguistic structure. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 6649–6663. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.367. URL https://doi.org/10.18653/v1/2023.acl-long.367.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Pei Chen, Haotian Xu, Cheng Zhang, and Ruihong Huang. Crossroads, buildings and neighborhoods: A dataset for fine-grained location recognition. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3329–3339, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.243. URL https://aclanthology.org/2022.naacl-main.243.

Yi Chen, Haiyun Jiang, Lemao Liu, Shuming Shi, Chuang Fan, Min Yang, and Ruifeng Xu. An empirical study on multiple information sources for zero-shot fine-grained entity typing. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2668–2678, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.210. URL https://aclanthology.org/2021.emnlp-main.210.

Zekai Chen, Mariann Micsinai Balan, and Kevin Brown. Language models are few-shot learners for prognostic prediction. *CoRR*, abs/2302.12692, 2023. doi: 10.48550/arXiv.2302.12692. URL https://doi.org/10.48550/arXiv.2302.12692.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022. doi: 10.48550/arXiv.2204.02311. URL https://doi.org/10.48550/arXiv.2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022. doi: 10.48550/arXiv.2210.11416. URL https://doi.org/10.48550/arXiv.2210.11416.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 8440–8451. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.747. URL https://doi.org/10.18653/v1/2020.acl-main.747.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. Broad Twitter corpus: A diverse named entity recognition resource. In Yuji Matsumoto and Rashmi Prasad (eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1169–1179, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1111.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. Results of the WNUT2017 shared task on novel and emerging entity recognition. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin (eds.), *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 140–147, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4418. URL https://aclanthology.org/W17-4418.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *CoRR*, abs/2305.14314, 2023. doi: 10.48550/arXiv.2305.14314. URL https://doi.org/10.48550/arXiv.2305.14314.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1286–1305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.98. URL https://aclanthology.org/2021.emnlp-main.98.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. Multi-sentence argument linking. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8057–8077, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.718. URL https://aclanthology.org/2020.acl-main.718.

Hermenegildo Fabregat, Juan Martínez-Romo, and Lourdes Araujo. Overview of the DIANN task: Disability annotation task. In Paolo Rosso, Julio Gonzalo, Raquel Martínez, Soto Montalvo, and Jorge Carrillo de Albornoz (eds.), *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pp. 1–14. CEUR-WS.org, 2018. URL https://ceur-ws.org/Vol-2150/overview-diann-task.pdf.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. Semeval-2023 task 2: Fine-grained multilingual named entity recognition (multiconer 2). In Atul Kr. Ojha, A. Seza Dogruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori (eds.), *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, pp. 2247–2265. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.semeval-1.310. URL https://doi.org/10.18653/v1/2023.semeval-1.310.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=sE7-XhLxHA.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Rezarta Islamaj Doğan and Zhiyong Lu. An improved corpus of disease mentions in PubMed citations. In Kevin B. Cohen, Dina Demner-Fushman, Sophia Ananiadou, Bonnie Webber, Jun'ichi Tsujii, and John Pestian (eds.), *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pp. 91–99, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://aclanthology.org/W12-2411.

Veysel Kocaman and David Talby. Biomedical named entity recognition at scale. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani (eds.), *Pattern Recognition. ICPR International Workshops and Challenges*, pp. 635–646, Cham, 2021. Springer International Publishing. ISBN 978-3-030-68763-2.

Aman Kumar and Binil Starly. "fabner": information extraction from manufacturing process science domain literature using named entity recognition. *J. Intell. Manuf.*, 33(8):2393–2407, 2022. doi: 10.1007/s10845-021-01807-x. URL https://doi.org/10.1007/s10845-021-01807-x.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In Roger Levy and Lucia Specia (eds.), *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 333–342, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-1034. URL https://aclanthology.org/K17-1034.

Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. CodeIE: Large code generation models are better few-shot information extractors. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15339–15353, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.855. URL https://aclanthology.org/2023.acl-long.855.

Sha Li, Heng Ji, and Jiawei Han. Document-level event argument extraction by conditional generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 894–908, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.69. URL https://aclanthology.org/2021.naacl-main.69.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. A joint neural model for information extraction with global features. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7999–8009, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.713. URL `https://aclanthology.org/2020.acl-main.713`.

Xiao Ling and Daniel S. Weld. Fine-grained entity recognition. In Jörg Hoffmann and Bart Selman (eds.), *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada*, pp. 94–100. AAAI Press, 2012. doi: 10.1609/AAAI.V26I1.8122. URL `https://doi.org/10.1609/aaai.v26i1.8122`.

Jingjing Liu, Panupong Pasupat, Scott Cyphers, and James R. Glass. Asgard: A portable architecture for multilingual dialogue systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pp. 8386–8390. IEEE, 2013. doi: 10.1109/ICASSP.2013.6639301. URL `https://doi.org/10.1109/ICASSP.2013.6639301`.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. Crossner: Evaluating cross-domain named entity recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 13452–13460. AAAI Press, 2021a. URL `https://ojs.aaai.org/index.php/AAAI/article/view/17587`.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. Crossner: Evaluating cross-domain named entity recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 13452–13460. AAAI Press, 2021b. doi: 10.1609/aaai.v35i15.17587. URL `https://doi.org/10.1609/aaai.v35i15.17587`.

Jie Lou, Yaojie Lu, Dai Dai, Wei Jia, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. Universal information extraction as unified semantic matching. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 13318–13326. AAAI Press, 2023. doi: 10.1609/aaai.v37i11.26563. URL `https://doi.org/10.1609/aaai.v37i11.26563`.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. Unified structure generation for universal information extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 5755–5772. Association for Computational Linguistics, 2022a. doi: 10.18653/v1/2022.acl-long.395. URL `https://doi.org/10.18653/v1/2022.acl-long.395`.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. Unified structure generation for universal information extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5755–5772, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.395. URL `https://aclanthology.org/2022.acl-long.395`.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 157–165, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.18. URL `https://aclanthology.org/2022.acl-short.18`.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. The E3C project: European clinical case corpus. In Jon Alkorta, Itziar Gonzalez-Dios, Aitziber Atutxa, Koldo Gojenola, Eugenio Martínez-Cámara, Álvaro Rodrigo, and Paloma Martínez (eds.), *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations (SEPLN-PD 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), Málaga, Spain, September, 2021*, volume 2968 of *CEUR Workshop Proceedings*, pp. 17–20. CEUR-WS.org, 2021. URL `https://ceur-ws.org/Vol-2968/paper5.pdf`.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2), sep 2023. ISSN 0360-0300. doi: 10.1145/3605943. URL `https://doi.org/10.1145/3605943`.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 15991–16111. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.891. URL `https://doi.org/10.18653/v1/2023.acl-long.891`.

Rasha Obeidat, Xiaoli Fern, Hamed Shahbazi, and Prasad Tadepalli. Description-based zero-shot fine-grained entity typing. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 807–814, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1087. URL `https://aclanthology.org/N19-1087`.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon LLM: outperforming curated corpora with web data, and web data only. *CoRR*, abs/2306.01116, 2023. doi: 10.48550/arXiv.2306.01116. URL `https://doi.org/10.48550/arXiv.2306.01116`.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards robust linguistic analysis using OntoNotes. In Julia Hockenmaier and Sebastian Riedel (eds.), *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 143–152, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `https://aclanthology.org/W13-3516`.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL `https://openreview.net/forum?id=R8sQPpGCv0`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In Suzanne Stevenson and Xavier Carreras (eds.), *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pp. 147–155, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL `https://aclanthology.org/W09-1119`.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton-Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code. *CoRR*, abs/2308.12950, 2023. doi: 10.48550/arXiv.2308.12950. URL https://doi.org/10.48550/arXiv.2308.12950.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. Label verbalization and entailment for effective zero and few-shot relation extraction. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1199–1212, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.92. URL https://aclanthology.org/2021.emnlp-main.92.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 2439–2455, Seattle, United States, July 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.187. URL https://aclanthology.org/2022.findings-naacl.187.

Oscar Sainz, Haoling Qiu, Oier Lopez de Lacalle, Eneko Agirre, and Bonan Min. ZS4IE: A toolkit for zero-shot information extraction with simple verbalizations. In Hannaneh Hajishirzi, Qiang Ning, and Avi Sil (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pp. 27–38, Hybrid: Seattle, Washington + Online, July 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-demo.4. URL https://aclanthology.org/2022.naacl-demo.4.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.722. URL https://aclanthology.org/2023.findings-emnlp.722.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, and Eneko Agirre. Did chatgpt cheat on your test?, Jun 2023b. URL https://hitz-zentroa.github.io/lm-contamination/blog/.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. Casie: Extracting cybersecurity event information from text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8749–8757, Apr. 2020. doi: 10.1609/aaai.v34i05.6401. URL https://ojs.aaai.org/index.php/AAAI/article/view/6401.

Teven Le Scao and Alexander M. Rush. How many data points is a prompt worth? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 2627–2636. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.208. URL https://doi.org/10.18653/v1/2021.naacl-main.208.

Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 255–269. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.20. URL https://doi.org/10.18653/v1/2021.eacl-main.20.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *CoRR*, abs/2212.13138, 2022. doi: 10.48550/ARXIV.2212.13138. URL https://doi.org/10.48550/arXiv.2212.13138.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, pp. 1–9, 2023.

Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, and Fei Huang. DAMO-NLP at semeval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition. In Atul Kr. Ojha, A. Seza Dogruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori (eds.), *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, pp. 2014–2028. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.semeval-1.277. URL https://doi.org/10.18653/v1/2023.semeval-1.277.

Simone Tedeschi and Roberto Navigli. MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 801–812, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.60. URL https://aclanthology.org/2022.findings-naacl.60.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003. URL https://aclanthology.org/W03-0419.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023a. doi: 10.48550/arXiv.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023b. doi: 10.48550/arXiv.2307.09288. URL https://doi.org/10.48550/arXiv.2307.09288.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45, 2006. URL https://catalog.ldc.upenn.edu/LDC2006T06.

Ben Wang and Aran Komatsuzaki. Gpt-j-6b: A 6 billion parameter autoregressive language model, 2021.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. Instructuie: Multi-task instruction tuning for unified information extraction. *CoRR*, abs/2304.08085, 2023a. doi: 10.48550/arXiv.2304.08085. URL https://doi.org/10.48550/arXiv.2304.08085.

Xingyao Wang, Sha Li, and Heng Ji. Code4Struct: Code generation for few-shot event structure prediction. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3640–3663, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.202. URL https://aclanthology.org/2023.acl-long.202.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. Improving named entity recognition by external context retrieving and cooperative learning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1800–1812, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.142. URL https://aclanthology.org/2021.acl-long.142.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 5085–5109. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.340. URL https://doi.org/10.18653/v1/2022.emnlp-main.340.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. *Database*, 2016, 2016.

Renzo M. Rivera Zabala, Paloma Martinez, and Isabel Segura-Bedmar. A hybrid bi-lstm-crf model to recognition of disabilities from biomedical texts. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages*, 2018. URL https://ceur-ws.org/Vol-2150/DIANN_paper5.pdf.

Mozhi Zhang, Hang Yan, Yaqian Zhou, and Xipeng Qiu. Promptner: A prompting method for few-shot named entity recognition via k nearest neighbor search. *CoRR*, abs/2305.12217, 2023a. doi: 10.48550/arXiv.2305.12217. URL https://doi.org/10.48550/arXiv.2305.12217.

Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. Optimizing bi-encoder for named entity recognition via contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=9EAQVEINuum.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068, 2022. doi: 10.48550/arXiv.2205.01068. URL https://doi.org/10.48550/arXiv.2205.01068.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 35–45, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1004. URL https://aclanthology.org/D17-1004.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685, 2023. doi: 10.48550/arXiv.2306.05685. URL `https://doi.org/10.48550/arXiv.2306.05685`.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. Universalner: Targeted distillation from large language models for open named entity recognition. *CoRR*, abs/2308.03279, 2023. doi: 10.48550/arXiv.2308.03279. URL `https://doi.org/10.48550/arXiv.2308.03279`.

```
@dataclass
class Launcher(Template):
    """Refers to a vehicle designed primarily to transport payloads from the Earth's
    surface to space. Launchers can carry various payloads, including satellites,
    crewed spacecraft, and cargo, into various orbits or even beyond Earth's orbit.
    They are usually multi-stage vehicles that use rocket engines for propulsion."""

    mention: str
    """
    The name of the launcher vehicle.
    Such as: "Sturn V", "Atlas V", "Soyuz", "Ariane 5"
    """
    space_company: str # The company that operates the launcher. Such as: "Blue origin", "ESA", "Boeing"
    crew: List[str]
    """Names of the crew members boarding the Launcher.
    Such as: "Neil Armstrong", "Michael Collins", "Buzz Aldrin"
    """

@dataclass
class Mission(Template):
    """Any planned or accomplished journey beyond Earth's atmosphere with specific objectives,
    either crewed or uncrewed. It includes missions to satellites, the International
    Space Station (ISS), other celestial bodies, and deep space."""

    mention: str
    """
    The name of the mission.
    Such as: "Apollo 11", "Artemis", "Mercury"
    """
    date: str # The start date of the mission
    departure: str # The place from which the vehicle will be launched. Such as: "Florida", "Houston"
    destination: str # The place or planet to which the launcher will be sent. Such as "Moon", "low-orbit"


# This is the text to analyze
text = (
    "The Ares 3 mission to Mars is scheduled for 2032. The Starship rocket build by SpaceX will take off"
    "from Boca Chica, carrying the astronauts Max Rutherford, Elena Soto, and Jake Martinez."
)

# The annotation instances that take place in the text above are listed here
result = [
    Mission(mention='Ares 3', date='2032', departure='Boca Chica', destination='Mars'),
    Launcher(mention='Starship', space_company='SpaceX', crew=['Max Rutherford', 'Elena Soto', 'Jake Martinez'])
]
```

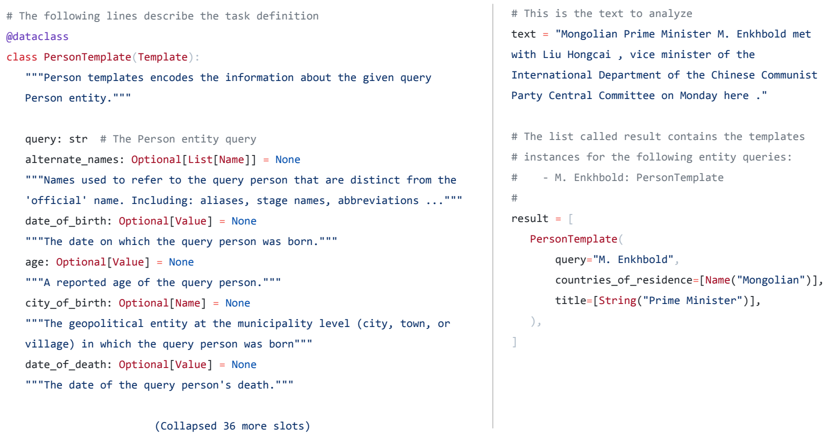
Figure 5: Example of generalization to custom tasks defined by the user.

A EXAMPLES

In addition to NER, EE and EAE for which examples are shown in Figures 2 and 3 respectively, we also feed the model with data from RE and SF. The formulation of RE is similar to the NER but with two argument attributes. However, the SF task is more complex as shown in the Figure 6. With this task, we added several layers of complexity to the input: extended definitions for each possible attribute (slot), optional arguments, and, fine-grained definitions of types such as Names, Values or Strings. We also added constraints into the prompt to condition the model to just output the information of our desired query instead of every single template on the text. In the future, we would like to add more complex tasks into the train and evaluation to improve the capabilities and flexibility of the model. For more examples refer to the GitHub repository.

A.1 EXAMPLE OF GENERALIZATION TO NEW CUSTOM TASKS

Our model allows the user to define custom annotation schemas using Python code. We provide an example where we define two new types of entities: Launcher and Mission. As shown in Figure 5, Launcher and Mission are not simple entities, they correspond to what we call *Template*, a class similar to *Entity* but with additional arguments, like the SF task. For example, the space_company or the crew of the launcher are some of the additional arguments we added to the schema. As shown in the example, the model's output (everything after `result = []`) satisfies the type constraints defined in the guidelines, attributes defined as strings are filled with strings and, the arguments defined as lists (like *crew*) are filled with lists. The model is able to correctly analyze the given sentence with our newly created annotation schema.



```
# The following lines describe the task definition
@dataclass
class PersonTemplate(Template):
    """Person templates encodes the information about the given query
    Person entity."""

    query: str  # The Person entity query
    alternate_names: Optional[List[Name]] = None
    """Names used to refer to the query person that are distinct from the
    'official' name. Including: aliases, stage names, abbreviations ..."""
    date_of_birth: Optional[Value] = None
    """The date on which the query person was born."""
    age: Optional[Value] = None
    """A reported age of the query person."""
    city_of_birth: Optional[Name] = None
    """The geopolitical entity at the municipality level (city, town, or
    village) in which the query person was born"""
    date_of_death: Optional[Value] = None
    """The date of the query person's death."""

                      (Collapsed 36 more slots)
```

```
# This is the text to analyze
text = "Mongolian Prime Minister M. Enkhbold met
with Liu Hongcai , vice minister of the
International Department of the Chinese Communist
Party Central Committee on Monday here ."

# The list called result contains the templates
# instances for the following entity queries:
#    - M. Enkhbold: PersonTemplate
#
result = [
   PersonTemplate(
       query="M. Enkhbold",
       countries_of_residence=[Name("Mongolian")],
       title=[String("Prime Minister")],
   ),
]
```

Figure 6: Example of the TACRED dataset converted to Slot Filling task represented as code.

B Performance in Seen vs Unseen Labels: Further Analysis

Table 6: List of labels in the zero-shot datasets that overlap with the ones in the training datasets (seen) and the labels that do not overlap with the ones in the training datasets (unseen)

Dataset	Seen Labels	Unseen Labels
BroadTwitter	Location, Organization, Person	-
CASIE$_{EE}$	-	DatabreachAttack, PhisingAttack, RansomAttack, VulnerabilityDiscover, VulnerabilityPatch
AI	Product, Country, Person, Organization, Location, Miscellaneous	Field, Task, Algorithm, Researcher, Metric, University, ProgrammingLanguage, Conference
Literature	Event, Person, Location, Organization, Country, Miscellaneous	Book, Writer, Award, Poem, Magazine, LiteraryGenre
Music	Event, Country, Location, Organization, Person, Miscellaneous	MusicGenre, Song, Band, Album, MusicalArtist, MusicalInstrument, Award
Politics	Person, Organization, Location, Election, Event, Country, Miscellaneous	Politician, PoliticalParty
Science	Person, Organization, Country, Location, ChemicalElement, ChemicalCompound, Event, Miscellaneous	Scientist, University, Discipline, Enzyme, Protein, AstronomicalObject, AcademicJournal, Theory, Award
E3C	ClinicalEntity	-
FabNER	Biomedical	Material, ManufacturingProcess, MachineEquipment, Application, EngineeringFeatures, MechanicalProperties, ProcessCharacterization, ProcessParameters, EnablingTechnology, ConceptPrinciples, ManufacturingStandards
HarveyNER	-	Point, Area, Road, River
Movie	Year	Actor, Character, Director, Genre, Plot, Rating, RatingsAverage, Review, Song, Tittle, Trailer
Restaurants	Location, Price, Hours	Rating, Amenity, RestaurantName, Dish, Cuisine
MultiNERD	Person, Location, Organization, Biological, Disease, Event, Time, Vehicle	Animal, Celestial, Food, Instrument, Media, Plant, Mythological
WikiEvents$_{NER}$	CommercialProduct, Facility, GPE, Location, MedicalHealthIssue, Money, Organization, Person, JobTitle, Numeric, Vehicle, Weapon	Abstract, BodyPart, Information, SideOfConflict
WikiEvents$_{EE}$	ConflictEvent, ContactEvent, GenericCrimeEvent, JusticeEvent, MedicalEvent, MovementTransportEvent, PersonnelEvent, TransactionEvent	ArtifactExistanceEvent, CognitiveEvent, ControlEvent, DisasterEvent, LifeEvent

Table 6 categorizes the labels for each zero-shot dataset into those that overlap with the training dataset and those that are completely unseen. We adhere to a strict approach in this classification. For instance, although the label COUNTRY does not appear in the training datasets, similar labels such as GEOPOLITICAL entity do. Therefore, we consider that the model has been exposed to this label during training.

While some labels in the zero-shot datasets overlap with those in the training dataset, the annotation guidelines for each label may vary significantly between datasets. Table 7 presents the micro-F1

Table 7: Micro F1 score for the seen and unseen labels in the zero-shot datasets.

	Baseline		🐶		🐶 13B		🐶 34B	
Dataset	**Seen**	**Unseen**	**Seen**	**Unseen**	**Seen**	**Unseen**	**Seen**	**Unseen**
BroadTwitter	39.0$_{\pm0.6}$	-	49.5$_{\pm0.8}$	-	51.4$_{\pm1.8}$	-	50.3$_{\pm2.1}$	-
CASIE$_{\text{EE}}$	-	33.9$_{\pm6.5}$	-	59.3$_{\pm2.3}$	-	62.2$_{\pm0.9}$	-	65.5$_{\pm1.8}$
AI	43.5$_{\pm1.4}$	21.1$_{\pm0.3}$	57.8$_{\pm1.2}$	60.0$_{\pm1.2}$	57.8$_{\pm0.8}$	55.8$_{\pm4.7}$	57.7$_{\pm2.8}$	64.2$_{\pm1.3}$
Literature	34.6$_{\pm0.2}$	43.6$_{\pm1.5}$	54.6$_{\pm3.6}$	67.4$_{\pm3.0}$	52.4$_{\pm0.2}$	64.6$_{\pm0.5}$	52.7$_{\pm2.1}$	63.7$_{\pm2.8}$
Music	46.8$_{\pm1.0}$	62.2$_{\pm1.6}$	53.7$_{\pm0.2}$	74.9$_{\pm0.3}$	52.8$_{\pm3.9}$	72.7$_{\pm3.5}$	54.0$_{\pm3.8}$	76.3$_{\pm1.2}$
Politics	45.9$_{\pm1.1}$	4.6$_{\pm2.6}$	64.0$_{\pm0.2}$	31.9$_{\pm4.7}$	62.0$_{\pm2.2}$	22.4$_{\pm14.6}$	64.4$_{\pm1.5}$	45.8$_{\pm9.3}$
Science	38.7$_{\pm0.8}$	34.7$_{\pm3.0}$	52.7$_{\pm1.7}$	58.8$_{\pm1.5}$	52.7$_{\pm1.0}$	60.4$_{\pm0.9}$	52.5$_{\pm0.4}$	60.5$_{\pm0.7}$
E3C	59.8$_{\pm0.3}$	-	59.0$_{\pm0.7}$	-	59.0$_{\pm0.9}$	-	60.0$_{\pm0.4}$	-
FabNER	0.0$_{\pm0.0}$	6.2$_{\pm0.4}$	22.6$_{\pm2.3}$	24.9$_{\pm0.6}$	23.9$_{\pm4.4}$	25.5$_{\pm0.6}$	20.7$_{\pm2.9}$	26.5$_{\pm0.6}$
HarveyNER	-	23.2$_{\pm0.4}$	-	37.3$_{\pm1.8}$	-	41.3$_{\pm0.9}$	-	38.9$_{\pm0.5}$
Movie	31.5$_{\pm0.7}$	46.1$_{\pm1.5}$	58.7$_{\pm2.3}$	63.8$_{\pm0.5}$	47.3$_{\pm3.1}$	65.3$_{\pm1.0}$	42.7$_{\pm2.3}$	66.1$_{\pm1.4}$
Restaurants	18.0$_{\pm1.1}$	38.7$_{\pm2.8}$	33.2$_{\pm2.7}$	49.9$_{\pm1.5}$	38.0$_{\pm3.6}$	57.1$_{\pm0.2}$	46.0$_{\pm4.2}$	57.2$_{\pm0.9}$
MultiNERD	58.0$_{\pm1.1}$	39.5$_{\pm1.4}$	81.2$_{\pm0.5}$	44.6$_{\pm0.9}$	82.4$_{\pm0.4}$	47.7$_{\pm0.7}$	82.3$_{\pm0.5}$	49.1$_{\pm0.5}$
WikiEvents$_{\text{NER}}$	77.2$_{\pm5.1}$	0.0$_{\pm0.0}$	81.5$_{\pm0.7}$	0.0$_{\pm0.0}$	80.9$_{\pm0.8}$	0.0$_{\pm0.0}$	82.1$_{\pm0.5}$	3.5$_{\pm2.6}$
WikiEvents$_{\text{EE}}$	43.3$_{\pm0.3}$	57.2$_{\pm1.5}$	41.7$_{\pm0.1}$	45.0$_{\pm1.5}$	43.9$_{\pm0.8}$	48.8$_{\pm1.7}$	45.0$_{\pm1.1}$	50.4$_{\pm0.9}$
Average	41.2$_{\pm0.4}$	31.6$_{\pm0.6}$	54.6$_{\pm0.3}$	47.5$_{\pm0.6}$	54.2$_{\pm0.4}$	48.0$_{\pm1.3}$	54.7$_{\pm0.9}$	51.4$_{\pm1.0}$

scores for both seen and unseen labels across each zero-shot dataset. Generally, the models perform better on seen labels compared to unseen ones. However, there are instances where the reverse is true. This occurs when a dataset contains labels that, although overlapping with those in the training dataset, have vastly different annotation guidelines. As discussed in Section 5.3, the model has strong preconceptions for some labels, adversely affecting zero-shot performance. GoLLIE, trained to adhere to specific annotation guidelines, demonstrates greater robustness against these label preconceptions than the baseline model. Consequently, it achieves better results for both seen and unseen labels. GoLLIE can successfully handle both, seen and unseen labels from datasets that were not used during training. This ability underscores GoLLIE's superior generalization capabilities, largely attributable to its capability of leveraging annotation guidelines.

C Model hallucinations

Table 8: Number impossible to parse outputs and number predicted labels that are hallucinations. F1 scores on the dataset are shown for reference.

	🐶		
Dataset	**Impossible to Parse**	**Hallucinations**	**F1 Score**
BroadTwitter	0$_{\pm0}$ / 2002	0$_{\pm0}$ / 1664	49.5$_{\pm0.8}$
CASIE$_{\text{EE}}$	1$_{\pm0}$ / 199	6$_{\pm1}$ / 1548	59.3$_{\pm2.3}$
CASIE$_{\text{EAE}}$	1$_{\pm1}$ / 199	3$_{\pm2}$ / 2804	50.0$_{\pm1.1}$
AI	0$_{\pm0}$ / 431	1$_{\pm1}$ / 1292	59.1$_{\pm1.1}$
Literature	0$_{\pm0}$ / 416	0$_{\pm0}$ / 2059	62.7$_{\pm3.2}$
Music	0$_{\pm0}$ / 465	6$_{\pm2}$ / 3080	67.8$_{\pm0.2}$
Politics	0$_{\pm0}$ / 651	3$_{\pm2}$ / 4142	57.2$_{\pm1.0}$
Science	0$_{\pm0}$ / 543	7$_{\pm1}$ / 2700	55.5$_{\pm1.6}$
E3C	0$_{\pm0}$ / 851	1$_{\pm0}$ / 688	59.0$_{\pm0.7}$
FabNER	1$_{\pm0}$ / 2064	13$_{\pm3}$ / 4474	24.8$_{\pm0.6}$
HarveyNER	0$_{\pm0}$ / 1303	1$_{\pm1}$ / 708	37.3$_{\pm1.8}$
Movie	0$_{\pm0}$ / 2443	1$_{\pm0}$ / 3919	63.0$_{\pm0.6}$
Restaurants	0$_{\pm0}$ / 1521	3$_{\pm0}$ / 1451	43.4$_{\pm0.8}$
MultiNERD	49$_{\pm11}$ / 32908	51$_{\pm8}$ / 67142	76.0$_{\pm0.7}$
WikiEvents$_{\text{NER}}$	0$_{\pm0}$ / 573	1$_{\pm1}$ / 2666	80.7$_{\pm0.7}$
WikiEvents$_{\text{EE}}$	0$_{\pm0}$ / 573	3$_{\pm1}$ / 630	43.0$_{\pm0.6}$
WikiEvents$_{\text{EAE}}$	2$_{\pm1}$ / 321	0$_{\pm0}$ / 363	51.9$_{\pm0.4}$

In this section, we evaluate the hallucinations generated by the model. We examine two different phenomena. First, we consider instances where the output is so corrupted that it is *impossible to parse*. In such cases, we treat the output as an empty list. Second, we look at instances where the model outputs a label *Hallucination*, that is, a label not defined among the input classes. In these instances, we remove the label from the output. As demonstrated in Table 8, for all the zero-shot datasets, both phenomena occur in less than 1% of the predictions. This demonstrates that GoLLIE is highly resistant to hallucinations and closely adheres to the classes defined in the input.

D Extended training details

D.1 Loss calculation

We have used the standard Next Token Prediction (NTP) loss to train our models. However, several regularizations that we applied to the models made the loss computed over the guideline tokens much higher than the actual output tokens. This is because we randomly shuffle the guidelines order, mask names or, drop classes, which makes impossible to predict what goes next. To avoid the loss of the guideline tokens overshadow the actual output tokens loss, we decided to only compute the loss over the output tokens. This way, we can also avoid some overfitting on the guidelines. This resulted on a faster training and better results overall.

Table 9: Number of examples for each training and zero-shot dataset

Dataset	**Train**	**Dev**	**Test**
ACE05$_{\text{NER}}$	19217	676	901
ACE05$_{\text{RE}}$	19217	901	676
ACE05$_{\text{EE}}$	19217	676	901
ACE05$_{\text{EAE}}$	3843	397	368
ACE05$_{\text{RC}}$	5691	-	-
ACE05$_{\text{VER}}$	19217	-	-
BC5CDR	4561	4582	4798
CoNLL 2003	14041	3250	3453
DIANN	3976	793	1309
NCBIDisease	5433	924	941
Ontonotes 5	*30000*	15680	12217
RAMS	7329	924	871
TACRED	10027	3896	2311
WNUT 2017	3394	1009	1287
Total	165163	33708	30033
BroadTwitter	-	-	2002
CASIE$_{\text{EE}}$	-	-	199
CASIE$_{\text{EAE}}$	-	-	199
AI	-	-	431
Literature	-	-	416
Music	-	-	465
Politics	-	-	651
Science	-	-	543
E3C	-	-	851
FabNER	-	-	2064
HarveyNER	-	-	1303
Movie	-	-	2443
Restaurants	-	-	1521
MultiNERD	-	-	32908
WikiEvents$_{\text{NER}}$	-	-	573
WikiEvents$_{\text{EE}}$	-	-	573
WikiEvents$_{\text{EAE}}$	-	-	321
Total	-	-	47463

D.2 Dataset Detailts

Table 9 shows the number of examples for each training and zero-shot dataset. OntoNotes is generated semi-automatically and is orders of magnitude larger than the other ones. Therefore, each training epoch, we sample 30.000 random examples from the training set. The models were trained for 3 epochs with an effective batch-size of 32 and a learning-rate of 3e-4 with cosine scheduler. Therefore, we perform 15,485 training steps.

Regarding the splits, we use the standard train, dev test splits for every dataset. In case of ACE, we follow the split provided by Lin et al. (2020). In the case of CASIE we took the first 200 instances as validation and the last 2000 as test.

Table 10: Details about the training resources required for each model.

Model	Hardware	FLOPs	Time (h)	CO_2eq (kg)
Baseline	1xA100	$4.5e^{18}$	17.3	0.61
GoLLIE	1xA100	$11.9e^{18}$	44.5	1.57
13B	1xA100	$22.7e^{18}$	79.5	2.80
34B	2xA100	$55.8e^{18}$	94.6	6.67

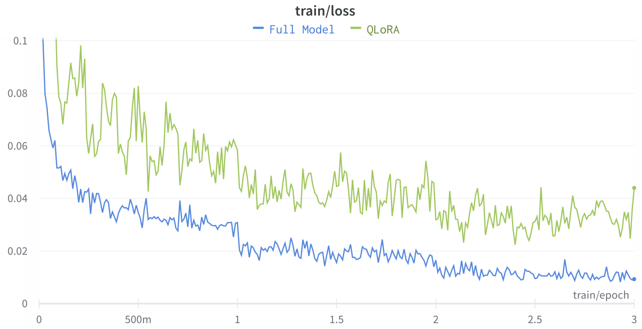
D.3 Carbon footprint

Fine-tuning LLMs is not as expensive as pre-training these models. Still, we believe that it is important to measure and document the costs that our experiments have on our planet. We provide the resources required for a single run of our experiments in Table 10. All the experiments were done on our private infrastructure. For the carbon footprint estimation, we estimated the values considering a 400W consumption per GPU with a 0.141 kg/kWh carbon intensity[†].

D.4 LoRA vs Full Model Fine-Tuning

We conducted preliminary experiments to compare the performance of QLoRA (Hu et al., 2022; Dettmers et al., 2023) with that of training all the parameters in the model. These preliminary experiments were conducted using the LLaMA2 7B model Touvron et al. (2023b) and an early version of the code. However, the experimental setup for both models was identical. Both approaches were prompted with guidelines. First, we compared the training loss of both approaches. Figure 7 shows that when fine-tuning all the parameters, the loss decreases much more rapidly than when training only the LoRA layers. It also achieves a lower loss at the end of training. However, when evaluating the model at the end of the first and third epochs, we observed that training the full model performs very poorly, as shown in Table 11. We hypothesize that when training all the parameters, the model overfits quickly (indicated by the lower training loss) and memorizes the training data. On the other hand, training only the LoRA layers, which represent around 0.5% of the total model weights, introduces a bottleneck that prevents the model from memorizing the training dataset. It is also noteworthy that the QLoRA approach was trained using just one Nvidia A100 80GB GPU thanks to 4 bits quantization of the frozen model Dettmers et al. (2023). Training the full model required a minimum of four Nvidia A100 80GB GPUs to fit the model into memory. We used Deep-Speed [‡] to distribute the model across the four GPUs for training. Due to the high cost of training, we did not perform an extensive hyper-parameter search for the full model.

Figure 7: Training loss of fine-tuning the full model vs training LoRA layers only.



[†] Statistic taken from `https://app.electricitymaps.com/map`

[‡] `github.com/microsoft/DeepSpeed`

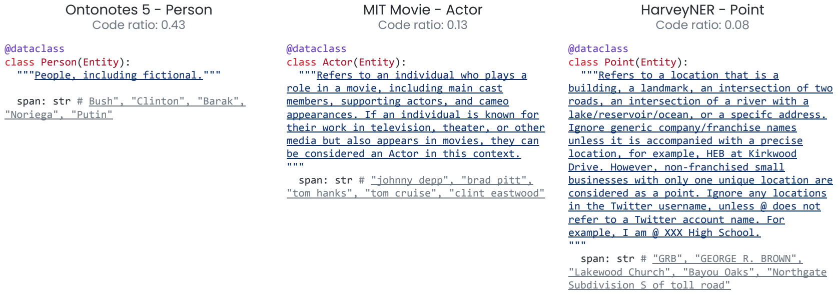
Table 11: F1 scores achieved when training the full model vs only training the LoRA Layers at the end of the first and third epoch.

Training	Epoch	Precision	LR	HarveyNer	FabNER	Restaurant	Movie	CASIE$_{EE}$	CoNLL03
Full	1	BF16	1e−4	0.00	0.00	0.25	4.74	0.00	85.57
Full	3	BF16	1e−4	3.45	0.21	**46.7**	16.72	0.42	84.83
QLoRA	1	4Bit + BF16	2e−3	34.98	**20.78**	45.01	**51.14**	55.83	91.41
QLoRA	3	4Bit + BF16	2e−3	**35.34**	16.21	39.07	44.18	**57.93**	**93.14**

E HANDLING DATASETS WITH HUNDREDS OF LABELS AND CODE-STYLE PROMPT OVERHEAD

In our research, we focus on datasets with fewer than 20 labels. However, some datasets, such as FIGER Ling & Weld (2012), include hundreds of fine-grained labels. Including guidelines in datasets with hundreds of labels can make inputs excessively long, exceeding the context size of current Large Language Models (LLMs). This is a known constraint in LLMs, and recently, significant research effort has been directed towards algorithms that efficiently increase the context window size Press et al. (2022). We anticipate that future LLMs will have a context window large enough to accommodate not only more labels but also more detailed guidelines. For the time being, this problem can be mitigated by batching the labels into multiple input examples. Instead of prompting the model with, for example, 100 labels in a single input, it is possible to prompt the model with 10 inputs, each incorporating 10 labels, and then combine all the outputs into a single response. In any case, handling datasets with a large number of labels remains a limitation of GoLLIE.

Figure 8: Percentage of characters from the input required to represent the code-style prompt for different labels. For detailed guidelines, the code is a small fraction of the input.



Ontonotes 5 - Person
Code ratio: 0.43

```
@dataclass
class Person(Entity):
  """People, including fictional."""

  span: str # Bush", "Clinton", "Barak",
"Noriega", "Putin"
```

MIT Movie - Actor
Code ratio: 0.13

```
@dataclass
class Actor(Entity):
  """Refers to an individual who plays a
role in a movie, including main cast
members, supporting actors, and cameo
appearances. If an individual is known for
their work in television, theater, or other
media but also appears in movies, they can
be considered an Actor in this context.
"""
  span: str # "johnny depp", "brad pitt",
"tom hanks", "tom cruise", "clint eastwood"
```

HarveyNER - Point
Code ratio: 0.08

```
@dataclass
class Point(Entity):
  """Refers to a location that is a
building, a landmark, an intersection of two
roads, an intersection of a river with a
lake/reservoir/ocean, or a specifc address.
Ignore generic company/franchise names
unless it is accompanied with a precise
location, for example, HEB at Kirkwood
Drive. However, non-franchised small
businesses with only one unique location are
considered as a point. Ignore any locations
in the Twitter username, unless @ does not
refer to a Twitter account name. For
example, I am @ XXX High School.
"""
  span: str # "GRB", "GEORGE R. BROWN",
"Lakewood Church", "Bayou Oaks", "Northgate
Subdivision S of toll road"
```

Our approach uses Python based code-style prompts, this requires to include tokens in the input to represent the code structures. Figure 8 illustrates various labels formatted in our code-style input alongside their respective guidelines. For very generic guidelines, such as those for the PERSON entity in OntoNotes 5, the code structure accounts for almost half of the input's characters. However, for detailed guidelines, like those for the POINT entity in HarveyNER, the code structure constitutes only a small portion of the input. While there is an overhead of tokens to represent the code structure, when dealing with datasets with a very large number of labels, the primary limitation is fitting the guideline definitions into the model's input, rather than accommodating the Python code structure.

F Human effort to build the prompts

GoLLIE requires formatting the input in a Python-based code representation. We achieve this by filling pre-defined templates for each task (NER, EE, EAE, RE, SF). We will make this predefined set of templates publicly available along with our code. Implementing new datasets only requires defining a list of labels and the guidelines for each label. We reuse the annotation guidelines provided by the dataset authors. Therefore, for most datasets, this process is straightforward and requires very little human effort. For datasets with very large and complex guidelines, such as TACRED, manual summarization of the guidelines was necessary. In any case, the involvement of a domain expert is not required. Additionally, since the inputs are automatically generated using templates, implementing new datasets does not require knowledge of Python coding.

Some datasets did not have publicly available guidelines, either due to semi-automatic generation or because the authors chose not to disclose them. For these specific datasets, human experts were needed to construct the guidelines from examples in the development split. We plan to release our generated guidelines to support future research. Human domain experts are necessary to adapt GoLLIE to new tasks where guidelines or annotations are unavailable. However, this requirement is common to any other Information Extraction (IE) model.

G Data-contamination statement

We believe data-contamination is a relevant problem that affects the NLP evaluations nowadays, becoming more prevalent with LLMs (Dodge et al., 2021; Magar & Schwartz, 2022; Sainz et al., 2023b;a). Detecting whether a dataset was inside a LLM pretrained corpora is challenging even with the pre-training data itself. In this paper, unfortunately, we do not have access to the pre-training data used to train Code-LLaMA the backbone LLM of our model. This issue is particularly worrying for us because one big source of contamination is probably GitHub and other code repositories which are also used to upload evaluation benchmarks Dodge et al. (2021). As Code-LLaMA is trained on code, there is a chance for this particular data-leakage. However, all of our comparisons were made against our baseline, which has the same backbone LLM as GoLLIE. Even if the results were impacted, **the improvements of our model over the baseline would not be affected by data-contamination as both share the same pre-training**. We hope that in the future more transparency will allow to perform safer evaluations.