

Ikerketa hizkuntza- ingeniaritzan

*Zer laguntza eman dezake makinak, euskaraz
argitaratzen den guztia eskura edukita?*

Arantza Díaz de Ilarraza Sánchez

Ixa Taldea

<http://ixa.eus>



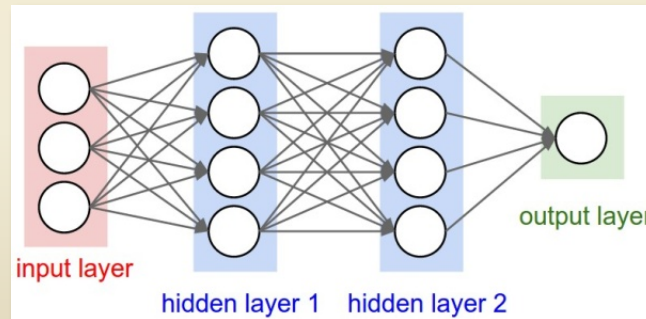
Big Data eta hizkuntza- teknologiak

- ◆ “Datu Handien” garaian bizi gara.
- ◆ Egunero hainbat trilioi hitz ekoizten dira.
- ◆ Urtez urte gero eta gehiago.
- ◆ Euskaraz ere bai.



Big Data eta hizkuntza- teknologiak

- ◆ Teknika berriak daude informazio hori baliatzeko:
 - Hodeiko konputazioa
 - Ikasketa sakona (*deep learning*)
 - Neurona-sareak





Big Data, hizkuntza-teknologiak eta euskara

- ◆ Baina teknika berri horiek...
baliagarriak al dira euskararako?
- ◆ ... testu kopurua ingelesezkoa baino askoz
txikiagoa izanik ere?





Edukiak

- ◆ Sarrera: Ixa Taldea
- ◆ Aplikazio-arloak
 - Testuen prozesaketa
 - Itzulpengintza automatikoa
 - Humanitate digitalak
 - Medikuntza
 - Hizkuntzen ikaskuntza





Ixa Taldea



2017ko ekaina



UPV/EHUko Ixa Taldea

- ◆ Hizkuntzaren tratamendu automatikoan aritzen den ikerketa-taldea (50 pertsona baino gehiago).
- ◆ Duela 30 urte sortua.
- ◆ Informatikariak eta hizkuntzalariak elkarlanean.
- ◆ Hizkuntzak: euskara, ingelesa, gaztelania...
- ◆ Lankidetzak: Elhuyar, Aholab, Tecnalía, Vicomtech, Langune, UEU, Iker (Baiona), UZEI... Microsoft, Google.
- ◆ Produktuak: zuzentzaileak, itzultzaileak, hiztegi elektronikoak, testu-corpusak...





Edukiak

- ◆ Sarrera: Ixa Taldea
- ◆ Aplikazio-arloak
 - **Testuen prozesaketa**
 - Itzulpengintza automatikoa
 - Humanitate digitalak
 - Medikuntza
 - Hizkuntzen ikaskuntza

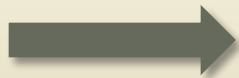


Testuen prozesaketa

◆ Informazioa erauzi

- Entitateak identifikatu: Carles Puigdemont, Mariano Rajoy, Madril, Katalunia... eta lotu baliabide eleaniztun eta zabalekin: Wikipedia eta bestelako baliabidekin.
- Gertaerak (*events*), denbora-adierazpenak... Identifikatu.

◆ Semantikan oinarritutako bilaketa aurreratuak ahalbideratzeko

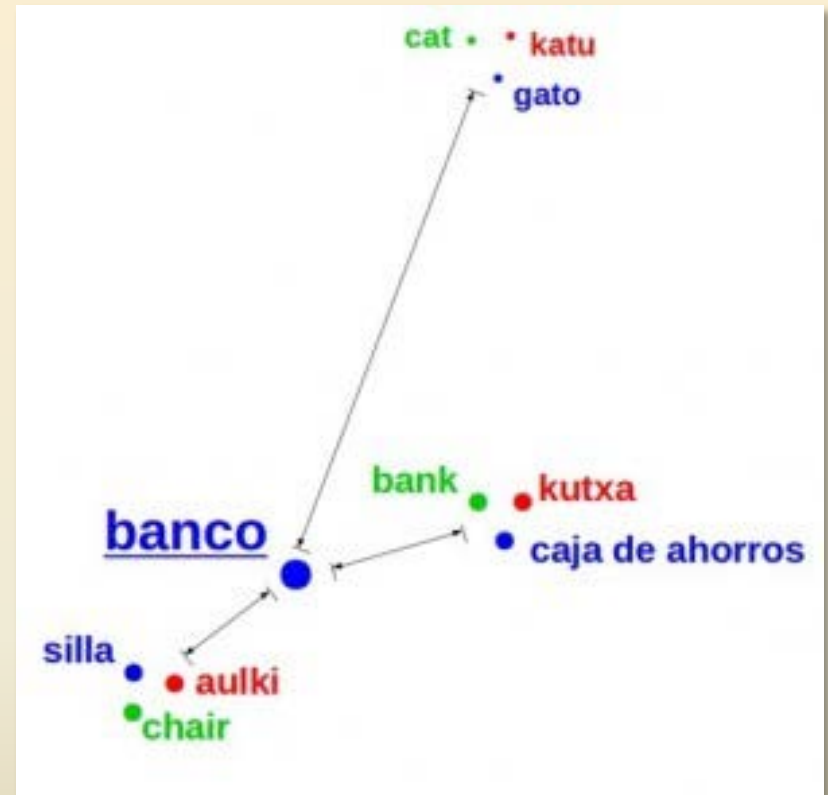


Informazioa errepresentatu behar



Testuen prozesaketa

- ◆ Hitzen esanahiak errepresentatu, beren arteko "distantziak" neurtu
- ◆ Zein dago gertu zeinetatik?



Testuen prozesaketa

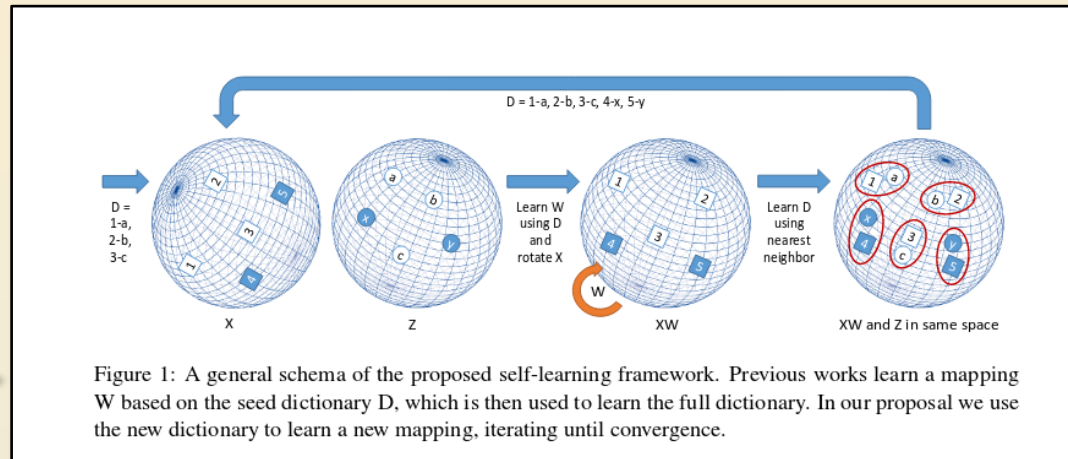
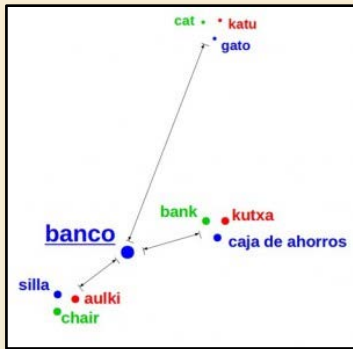


Figure 1: A general schema of the proposed self-learning framework. Previous works learn a mapping W based on the seed dictionary D , which is then used to learn the full dictionary. In our proposal we use the new dictionary to learn a new mapping, iterating until convergence.

Hitzen errepresentazio grafiko horrekin hiztegi elebidunak sor daitezke automatikoki (orokorrak edo espezializatuak).





Edukiak

- ◆ Sarrera: Ixa Taldea
- ◆ Aplikazio-arloak
 - Testuen prozesaketa
 - **Itzulpengintza automatikoa**
 - Humanitate digitalak
 - Medikuntza
 - Hizkuntzen ikaskuntza

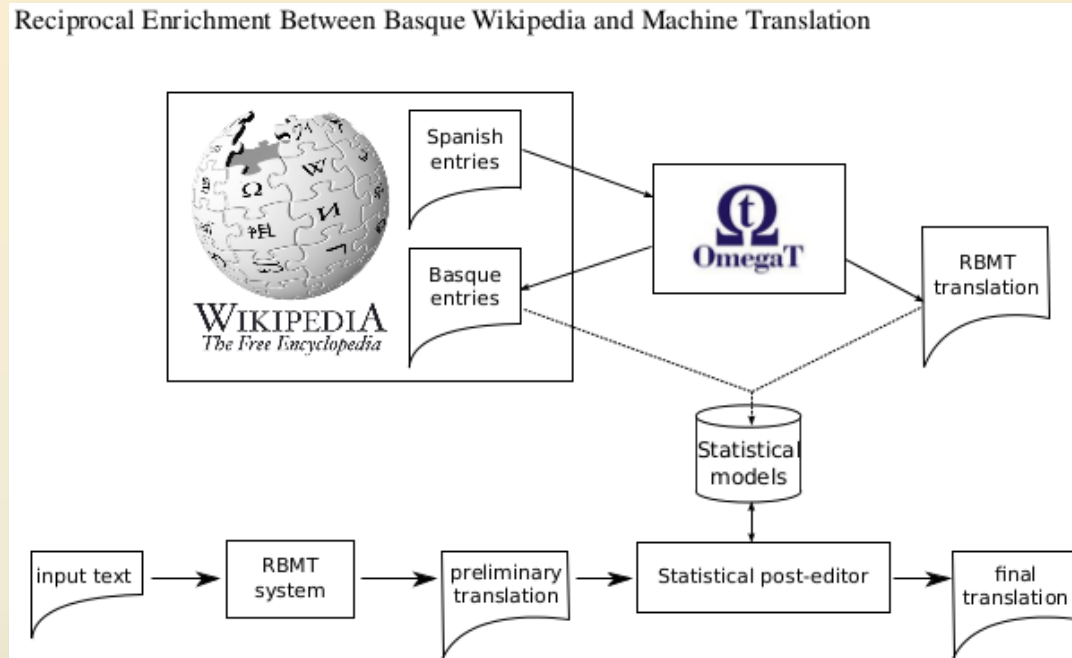


Itzulpengintza automatikoa

- ◆ Arlo interesgarria, ezagutzen ez ditugun hizkuntzetan idatzitako testuak ulertzen laguntzeko... Baina hainbat arazo daude:
 - a) Lexiko-aukeraketa (desegokia batzuetan)
 - b) Esaldiaren osagarrien ordena
 - c) Kolokazioak, adierazpen idiomatikoak
 - d) Gramatikaltasuna
 - e) ...
- ◆ Metodoak
 - Erregeletan oinarritutako hurbilpenetatik...
 - ◆ Datuetan oinarritutakoetara (corpus elebidun handiak)
 - ◆ Sare neuronaletara (ikasketa sakona)



Itzulpengintza automatikoa



- ◆ 2013: euskal Wikipediako 100 artikulua sortu ziren.
- ◆ % 10eko hobekuntza itzultzaile automatikoan.





Edukiak

- ◆ Sarrera: Ixa Taldea
- ◆ Aplikazio-arloak
 - Testuen prozesaketa
 - Itzulpengintza automatikoa
 - **Humanitate digitalak**
 - Medikuntza
 - Hizkuntzen ikaskuntza

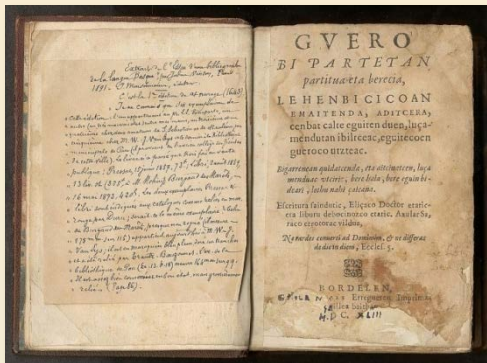


Humanitate digitalak

- ◆ Orain informatikariak eta hizkuntzalariak ari gara lankidetzan.
- ◆ Baina "ingurune digitalari" ekiteko, talde zabalagoak behar dira:
 - Historia
 - Kazetaritza
 - Soziologia
 - Psikologia
 - Zuzenbidea
 - ...
- ◆ Europan eta AEBn existitzen dira komunitate zabal horiek, lankidetzan ari direnak era naturalean.
- ◆ Estatu mailan, Ixak parte hartzen du Clarin azpiegitura-sarean (*Spanish Clarin Centre-K*).



Humanitate digitalak: testu historikoak



ikhusiagatik

Testu historikoak

Bilatu Zer da? Guri buruz Eguneratu indizeak Gehitu testua

Bilaketa: Mota:

17 emaitza, 3 testutan

Axular Gero ✕

har zuela eta halakoak bizia zor zuela, eta hala edekitzen zioten (Laert. lib.). Eta on lizate orai ere, halakoekin hala egin baledi.

Solon handiak ordenatu zuen, ezen aita batek bere semeari ofiziorik erakusten etzizanean, etzela seme hura bere aitaren faboratzera, beharrean **ikhusiagatik** ere, obligatu izanen (Plutarc. in Solone.). Zeren ofiziorik ez erakusteaz, alfer eta gaixto izateko bidean eta perilean utzi baitzen bere aitak.

Gimnosophista zerizten iende batzuek hain gaitzesten zuten alferkeria, ezen beñhere, afalaizinean deitzen baitzituzten presuna gazteak beregana, iakiteko ia zertan iragan zuten eguna, eta baldin frogatzen bazelen alferkeriarik, etzerauen afariak gaitzik egiten (Patricius lib. de republica).

Katon zensorino hartzaz irakurtzen da (zeñek baitzuen alferren gaitzen esku eta bothere) ekhartzen zeraukatenean gizon bat bere aitazinera, akusaturik, erraiten zela ezen alferra zela, berehala lehenbiziko gauza eskuetako larrua hazkatzen, eta ferekatzen zioela: eta baldin latz, lodi, eta gogor edireiten bazioen, ahalik eta arintkiena utzten zuela. Baiña mehe, leun eta bera bazuen, alferizat kondenaturik, falta guttiatik ere bortitzki gaztigitzen zuela (Plutarc. in vita Caton. Censor.).

Katon hark berak erraiten zuen, hirur gauzetarik, bere mendeen, ahal bezanbat, begi-

3* **BEHARDE LA**

har çuela eta halacoac bicia çorçuela, eta hala edequitcen çioten. Etá on liçate orai ere, halacoquin, hala eguin baledi.

Plutarc. in Solone. Solon handiac ordenatu çuen, ezen aita batek bere semeari, officioric eracufften etçioçnean, etçela seme hura, bere aitaren faboratzera, beharrean icçuffi agatic ere, obligatu izçanen. Çeren officioric ez eracuffteaz, alfer, eta gaixto içateico bidean eta perilean utçi baitçuen bere aitaç.

Patricius lib. de republica. Gimnosophista çerizten iende batçueç hain gaitz etçen çuten alferqueria, etç beñhere, afalaizinean deitcen baitçituzten presuna gazteac bere gana, içateico ea çortan iragan çuten eguna, eta baldin frogatcen baçeyen alferqueriaric, etçerauen afariac gaitçic eguiten.

Plutarc. in vita Caton. Censor. Caton Cenforino hartçaz iracurtçenda (etçieç baitçuen alferren gaincan eççu eta bothere) eçcartcen çeraucatenean guigoç bat bere aitaçinera, acçufaturic, erraiten eçla eçen alferra çela; berehala lehenbizico gauça eççuetaco larrua hazçatcen, eta fereçatcen eçioçla: eta baldin latz, lodi, eta gogor edireitcen baçioçen, ahalic eta arintçkiena utçzen çuela. Baiña baldin mehe, leun eta bera baçuen, alferizat condenaturic, falta gutti gatic ere bortitzqui gaztigitatcen çuela.

 Caton hare berac erraiten çuen, hirur gauzataric, bere mendeçan, ahal beçanbat, begi-

Iturria: Wikisource

2017-10-25

Euskarabildua 2017, Donostia

17

Testu historikoak, hizkuntza ez-normalizatu

- ◆ Sare sozialak
- ◆ Sentimentuen analisia: Behagunea proiektua

The screenshot displays the Behagunea web application interface. At the top, there is a navigation bar with the 'Behagunea' logo and a search bar. Below the navigation bar, there is a main content area with a word cloud on the left and a pie chart on the right. The word cloud contains terms like 'Euro-Dialogos', 'Biziz', 'Hirikilabs', 'Euro-Dialogues', 'Forum Theatre Ondarebideak', 'San Sebastian 2016', 'De_Ida_Y_Vuelta', and 'Hiruak'. The pie chart shows the distribution of sentiment: Positiboak (235, 77%), Negatiboak (47, 14%), and Neutroak (003, 1%). Below the word cloud and pie chart, there is a section for 'Menciones' with filters for 'Fecha: Hace un mes'. The mentions are listed in three columns: 'Todos', 'Menciones positivas', and 'Menciones negativas'. Each mention includes a tweet snippet and a timestamp.

Sentiment	Count	Percentage
Positiboak	235	77%
Negatiboak	47	14%
Neutroak	003	1%





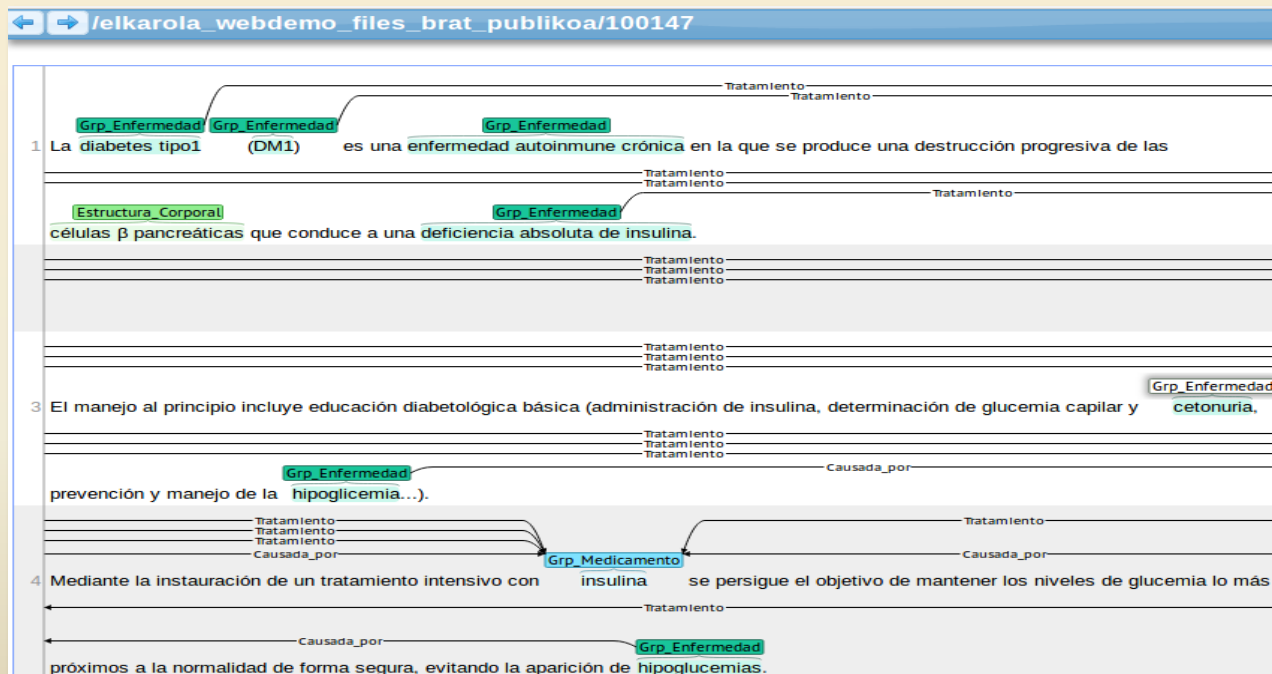
Edukiak

- ◆ Sarrera: Ixa Taldea
- ◆ Aplikazio-arloak
 - Testuen prozesaketa
 - Itzulpengintza automatikoa
 - Humanitate digitalak
 - **Medikuntza**
 - Hizkuntzen ikaskuntza



Medikuntza: testu medikoetatik informazioa erauzi

- ◆ Adibide bat: Botikek sortutako erreakzioak automatikoki identifikatzea osasun-txostenetan.



Medikuntza: testu medikoetatik informazioa erauzi

- ◆ Baina horretarako
- ◆ Osasun-txostenak euskaraz sortu.
- ◆ Terminologia finkatu behar da. Osasun-alorreko terminologia ez dago oraindik behar bezain landua.
- ◆ Ingelesezko 300.000 termino klinikotik gora ditu SNOMED CT datu-baseak; automatikoki euskaratua dago, eskuzko errebisioa falta da (Osakidetzan lantzen ari).





Edukiak

- ◆ Sarrera: Ixa Taldea
- ◆ Aplikazio-arloak
 - Testuen prozesaketa
 - Itzulpengintza automatikoa
 - Humanitate digitalak
 - Medikuntza
 - **Hizkuntzen ikaskuntza**





Hizkuntzen ikaskuntza

- ◆ Material didaktikoa eta ariketak sortzea, testu errealak erabilita.
- ◆ Idazlanen ebaluazioa: hiztegi-aberastasuna, espresioen erabilera, errore ortografiko eta sintaktikoak...
- ◆ Errore tipikoak identifikatzea. eta horiek zuzentzeko tresnak sortzea.
- ◆ Laburpenak egiten laguntzea.



Eskerrik asko!