IxaMed: Applying Freeling and a Perceptron Sequential Tagger at the Shared Task on Analyzing Clinical Texts

Koldo Gojenola, Maite Oronoz, Alicia Pérez, Arantza Casillas

IXA Taldea (UPV-EHU)
maite.oronoz@ehu.es
http://ixa.si.ehu.es

Abstract

This paper presents the results of the *IxaMed* team at the SemEval-2014 Shared Task 7 on Analyzing Clinical Texts. We have developed three different systems based on: a) exact match, b) a general-purpose morphosyntactic analyzer enriched with the SNOMED CT terminology content, and c) a perceptron sequential tagger based on a Global Linear Model. The three individual systems result in similar f-score while they vary in their precision and recall. We have also tried direct combinations of the individual systems, obtaining considerable improvements in performance.

1 Introduction

This paper presents the results of the *IxaMed* team. The task is focused on the identification (Task A) and normalization (Task B) of diseases and disorders in clinical reports.

We have developed three different systems based on: a) exact match, b) a general-purpose morphosyntactic analyzer enriched with the SNOMED CT terminology content, and c) a perceptron sequential tagger based on a Global Linear Model. The first system can be seen as a baseline that can be compared with other approaches, while the other two represent two alternative approaches based on knowledge organized in dictionaries/ontologies and machine learning, respectively. We also tried direct combinations of the individual systems, obtaining considerable improvements in performance.

These approaches are representative of different solutions that have been proposed in the literature (Pradhan et al., 2013), which can be broadly classified in the following types:

- *Knowledge-based.* This approach makes use of large-scale dictionaries and ontologies, that are sometimes integrated in general tools adapted to the clinical domain, as MetaMap (Aronson and Lang, 2010) and cTAKES (Xia et al., 2013).
- *Rule-based.* For example, in (Wang and Akella, 2013) the authors show the use of a rule-based approach on the output of MetaMap.
- *Statistical techniques.* These systems take a training set as input and apply different variants of machine learning, such as sequential taggers based on hidden Markov models (HMMs) or conditional random fields (CRFs) (Zuccon et al., 2013; Bodnari et al., 2013; Gung, 2013; Hervas et al., 2013; Leaman et al., 2013).
- *Combinations.* These approaches try to take the advantages of different system types, using methods such as voting or metaclassifiers (Liu et al., 2013).

In the rest of the paper, we will first introduce the different systems that we have developed in section 2, presenting the main results in section 3, and ending with the main conclusions.

2 System Description

The task of detecting diseases and their corresponding concept unique identifiers (CUI) has been faced using three methods that are described in the following subsections.

2.1 Exact Match

The system based on Exact Match (EM) simply obtained a list of terms and their corresponding CUI identifier from the training set and marked any appearance of those terms in the evaluation set. This simple method was improved with some additional extensions:

- *Improving precision.* In order to reduce the number of false positives (FP), we applied first the EM system to the training set itself. This process helped to measure FPs, for example, *blood* gave 184 FPs and 2 true positives (TPs). For the sake of not hurting the recall, we allowed the system to detect only those terms where $TP > FP$, that is, "blood" would not be classified as disorder.

- *Treatment of discontinuous terms.* For these terms, our system performed a soft-matching comparison allowing a limited variation for the text comprised between the term elements (for example "***right atrium** is mildly/moderately **dilated***"). These patterns were tuned manually.

2.2 Adapting Freeling to the Medical Domain

Freeling is an open-source multilingual language processing library providing a wide range of language analyzers for several languages (Padró et al., 2010), Spanish and English among others. We had already adapted Freeling to the medical domain in Spanish (Oronoz et al., 2013), so we used our previous experience to adapt the English version to the same domain. For the sake of clarity, we will refer to this system as FreeMed henceforth.

The linguistic resources (lexica, grammars,...) in Freeling can be modified, so we took advantage of this flexibility extending two standard Freeling dictionaries: a basic dictionary of terms consisting of a unique word, and a multiword-term dictionary. Both of them were enriched with a dictionary of medical abbreviations[1] and with the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) version dated 31st of July of 2013. In addition to the changes in the lexica, we added regular expressions in the tokenizer to recognize medical terms as *"Alzheimer's disease"* as a unique term.

In our approach, the system distinguishes between morphology and syntax on one side and semantics on the other side. First, on the morphosyntactic processing, our system only categorizes word-forms using their basic part-of-speech (POS) categories. Next, the semantic distinctions are applied (the identification of the term as substance, disorder, procedure,...). Following this approach, whenever the specific term on the new domain (biomedicine in this case) was already in Freeling's standard dictionaries, the specific entries will not be added to the lexicon. Instead, medical meanings are added in a later semantic tagging stage. For example: the widely used term *"fever"*, as common noun, was not added to the lexicon but its semantic class is given in a second stage. Only very specific terms not appearing in the lexica as, for instance, *"diskospondylitis"* were inserted. This solution helps to avoid an explosion of ambiguity in the morphosyntactic analysis and, besides, it enables a clear separation between morphosyntax and semantics.

In figure 1 the results of both levels of analysis, morphosyntactic and semantic, are shown. The linguistic and medical information of medical texts is stored in the Kyoto Annotation Format or KAF (Bosma et al., 2009) that is based in the eXtended Markup Language (XML). In this example the term *aneurysm* is analyzed as NN (meaning noun) and it is semantically categorized as *morphological abnormality* and *disorder*.

SNOMED CT is part of the Metathesaurus, one of the elements of the Unified Medical Language System (UMLS). We used the Metathesaurus vocabulary database to extract the mapping between SNOMED CT's concept identifiers and their corresponding UMLS's concept unique identifier (CUI). All the medical terms appearing in SNOMED CT and analyzed with FreeMed are tagged with both identifiers. For instance, the term *aneurysm* in figure 1 has the *85659009* SNOMED CT identifier when the term is classified in the *morphological abnormality* content hierarchy and the *432119003* identifier as *disorder*. Both are linked to the same concept identifier, *C0002940*, in UMLS. This mapping has been used for Task B, whenever the CUI is the same in all the analysis of the same term.

```
<term tid=”t241” lemma=”aneurysm” pos=”NN”>
 <extRefs>
  <extRef resource=”SCT_20130731” reference=”85659009”
     reftype=”morphologic_abnormality” >
    <extRef resource=”UMLS-2010AB” reference=”C0002940”/ >
  </extRef>
  <extRef resource=”SCT_20130731” reference=”432119003”
     reftype=”disorder” >
    <extRef resource=”UMLS-2010AB” reference=”C0002940”/>
  </extRef>
 </extRefs>
</term>
```

Figure 1: Analysis with augmented information.

All the terms from all the 19 content hierarchies of SNOMED CT were tagged with semantic information in the provided texts.

[1] http://www.jdmd.com/abbreviations-glossary.asp

The training corpus was linguistically analyzed and its format was changed from XML to the format specified at the shared task. After a manual inspection of the results and the Gold Standard, some selection and filtering of terms were performed:

- *Selection and combination of semantic classes.* All the terms from the *disorder* semantic class (for example *"Hypothyroidism"*) and from the *finding* class (for instance *"headache"* or *"dizziness"*) are chosen, as well as some tag combinations (see figure 1). After analyzing the train corpus we decided to join into a unique term a *body structure* immediately followed by a *disorder/finding*. In this way, we identify terms as *"MCA aneurysm"* that are composed of the *MCA* abbreviation (meaning the body structure *"middle cerebral artery"*) and the inmediately following *"aneurysm"* disorder.
- *Filtering of terms.* Not all the terms from the mentioned SNOMED CT hierarchies are identified as disorders in the Gold Standard. Some terms are discarded following these criteria: i) findings describing personal situations (e.g. *"alcoholic"*, *"lives with daughter"*,. . .), ii) findings describing current situations (e.g. *"awake"*,. . .), iii) findings with words indicating a negation or normal situation (e.g. *"stable blood pressure"*, *"no evidence of malignancy"*,. . .) and iv) too general terms (e.g. *"problems"*, *"illness"*,. . .).

The medical terms indicating disorders that are linked to more than one CUI identifier, were tagged as *CUI-less*. That is, we did not perform any CUI disambiguation.

In subsequent iterations and after analyzing our misses, new terms and term variations (Hina et al., 2013) are added to the lexica in Freeling with the restriction that, at least, one synonym should appear in SNOMED CT. Thus, equivalent forms were created for all the terms indicating a *cancer*, a *tumor*, a *syndrome*, or a specific *disease*. For instance, variants for the term *"cancer of colon"* and with the same SNOMED CT concept identifier (number 363406005) are created with the forms *"colon cancer"*, *"cancer of the colon"* and *"cancer in colon"*. Some abbreviation variations found in the Gold Standard are added in the lexica too, following the same criteria.

2.3 Perceptron Sequential Tagger

This system uses a Global Linear Model (GLM), a sequential tagger using the perceptron algorithm (Collins, 2002), that relies on Viterbi decoding of training examples combined with simple additive updates. The algorithm is competitive to other alternatives such as maximum-entropy taggers or CRFs.

The original textual files are firstly processed by FreeMed, and then the tagger uses all the available information to assign tags to the text. Each token contains information about the word form, lemma, part of speech, and SNOMED CT category.

Our GLM system only deals with Task A, and it will not tackle the problem of concept normalization, due to time constraints. In this respect, for Task B the GLM system will simply return the first SNOMED CT category given by FreeMed. This does not mean that GLM and FreeMed will give the same result for Task B, as the GLM system first categorizes each element as a disorder or disease, and it will gave a CUI only when that element has been identified.

2.4 Combinations

The previous subsections presented three different approaches to the problem that obtain comparable scores (see table 1). In the area of automatic tagging, there are several works that combine disparate systems, usually getting good results. For this reason, we tried the simplest approach of merging the outputs of the three individual systems into a single file.

3 Results

Table 1 presents the results of the individual and combined systems on the development set. Looking at the individual systems on Task A, we can see that all of them obtain a similar f-score, although there are important differences in terms of precision and recall. Contrary to our initial intuition, the FreeMed system, based on dictionaries and ontologies, gives the best precision and the lowest recall. In principle, having SNOMED CT as a base, we could expect that the coverage would be more complete (attaining the highest recall). However, the results show that there is a gap between the writing of the standard SNOMED CT terms and the terms written by doctors in their clinical notes. On the other hand, the sequential tagger gives the best recall. Since the tagger uses both

	Task A						Task B	
	Strict			Relaxed			Strict	Relaxed
System	**Precision**	**Recall**	**F-Score**	**Precision**	**Recall**	**F-Score**	**Accuracy**	
INDIVIDUAL SYSTEMS								
Exact Match (EM)	0.804	0.505	0.620	**0.958**	0.604	0.740	**0.479**	**0.948**
FreeMed	**0.822**	0.501	0.622	0.947	0.578	0.718	0.240	0.479
GLM	0.715	**0.570**	**0.634**	0.908	**0.735**	**0.813**	0.298	0.522
COMBINATIONS								
FreeMed + EM	**0.766**	0.652	**0.704**	**0.936**	0.754	0.835	**0.556**	0.855
FreeMed + GLM	0.689	0.668	0.678	0.903	0.790	0.843	0.345	0.518
EM + GLM	0.680	0.679	0.679	0.907	0.819	0.861	0.398	**0.598**
FreeMed + EM + GLM	0.659	**0.724**	0.690	0.899	**0.845**	**0.871**	0.421	0.584

Table 1: Results of the different systems on the development set.

	Task A						Task B	
	Strict			Relaxed			Strict	Relaxed
System	**Precision**	**Recall**	**F-Score**	**Precision**	**Recall**	**F-Score**	**Accuracy**	
FreeMed + EM	**0.729**	0.701	0.715	**0.885**	0.808	0.845	**0.604**	**0.862**
FreeMed + EM + GLM	0.681	**0.786**	**0.730**	0.872	**0.890**	**0.881**	0.439	0.558
Best system	**0.843**	**0.786**	**0.813**	**0.936**	**0.866**	**0.900**	**0.741**	**0.873**

Table 2: Results on the test set.

contextual words and prefixes and suffixes as features for learning, this method has proven helpful for the recognition of terms that do not appear in the training data (see the difference with respect to the exact match approach).

Looking at the different combinations in table 1, we see that two approaches work best, either combining FreeMed and EM, or combining the three individual systems. The inclusion of GLM results in the best coverage, but at the expense of precision. On the other hand, combining FreeMed and EM gives a better precision but lower coverage. As pointed out by Collins (2002), the results of the perceptron tagger are competitive with respect to other statistical approaches such as CRFs (Zuccon et al., 2013; Bodnari et al., 2013; Gung, 2013; Hervas et al., 2013; Leaman et al., 2013).

Regarding Task B, we can see that the EM system is by far the most accurate, while FreeMed is well below its a priori potential. The reason of this low result is mainly due to the high ambiguity found on the output of the SNOMED CT tagger, as many terms are associated with more than one CUI and, consequently, are left untagged. This problem deserves future work on automatic semantic disambiguation. On the combinations, FreeMed and EM together give the best result. However, as we told before, the GLM system was only trained for Task A, so it is not surprising to see that its results deteriorate the accuracy in Task B.

We chose these best two combinations for the evaluation on the test set (using training and development for experimentation or training), which are presented in table 2. Here we can see that results on the development also hold on the test set.

Given the unsophisticated approach to combine the systems, we can figure out more elaborated solutions, such as majority or weighted voting, or even more, the definition of a machine learning classifier to select the best system for every proposed term. These ideas are left for future work.

4 Conclusions

We have presented the IxaMed approach, composed of three systems that are based on exact match, linguistic and knowledge repositories, and a statistical tagger, respectively. The results of individual systems are comparable, with differences in precision and recall. We also tested a simple combination of the systems, which proved to give significant improvements over each individual system. The results are competitive, although still far from the winning system.

For future work, we plan to further improve the individual systems. Besides, we hope that the experimentation with new combination approaches will offer room for improvement.

Acknowledgements

This work was partially supported by the European Commission (325099 and SEP-210087649), the Spanish Ministry of Science and Innovation (TIN2012-38584-C06-02) and the Industry of the Basque Government (IT344-10).

References

Alan R Aronson and Francois-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association (JAMIA)*, 17:229–236.

Andreea Bodnari, Louise Deleger, Thomas Lavergne, Aurelie Neveol, and Pierre Zweigenbaum. 2013. A Supervised Named-Entity Extraction System for Medical Text. In *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop*, September.

Wauter Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini, and Carlo Aliprandi. 2009. KAF: a Generic Semantic Annotation Format. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon GL*, pages 17–19, Septembre.

Michael Collins. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 1–8. Association for Computational Linguistics, July.

James Gung. 2013. Using Relations for Identification and Normalization of Disorders: Team CLEAR in the ShARe/CLEF 2013 eHealth Evaluation Lab. In *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop*, September.

Lucia Hervas, Victor Martinez, Irene Sanchez, and Alberto Diaz. 2013. UCM at CLEF eHealth 2013 Shared Task1. In *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop*, September.

Saman Hina, Eric Atwell, and Owen Johnson. 2013. SnoMedTagger: A semantic tagger for medical narratives. In *Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*.

Robert Leaman, Ritu Khare, and Zhiyong Lu. 2013. NCBI at 2013 ShARe/CLEF eHealth Shared Task: Disorder Normalization in Clinical Notes with Dnorm. In *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop*, September.

Hongfang Liu, Kavishwar Wagholikar, Siddhartha Jonnalagadda, and Sunghwan Sohn. 2013. Integrated cTAKES for Concept Mention Detection and Normalization. In *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop*, September.

Maite Oronoz, Arantza Casillas, Koldo Gojenola, and Alicia Perez. 2013. Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names. In *Lecture Notes in Computer Science, 8259. Progress in Pattern Recognition, ImageAnalysis, ComputerVision, and Applications 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba*, November 20-23.

Lluis Padró, Samuel Reese, Eneko Agirre, and Aitor Soroa. 2010. Semantic Services in Freeling 2.1: WordNet and UKB. In *Global Wordnet Conference*, Mumbai, India.

Sameer Pradhan, Noemie Elhadad, Brett R. South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana Savova. 2013. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. In *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop*, September.

Chunye Wang and Ramakrishna Akella. 2013. UCSCs System for CLEF eHealth 2013 Task 1. In *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop*, September.

Yunqing Xia, Xiaoshi Zhong, Peng Liu, Cheng Tan, Sen Na, Qinan Hu, and Yaohai Huang. 2013. Combining MetaMap and cTAKES in Disorder Recognition: THCIB at CLEF eHealth Lab 2013 Task 1. In *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop*, September.

Guido Zuccon, Alexander Holloway, Bevan Koopman, and Anthony Nguyen. 2013. Identify Disorders in Health Records using Conditional Random Fields and Metamap AEHRC at ShARe/CLEF 2013 eHealth Evaluation Lab Task 1. In *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop*, September.