Multilingual Named-Entity Recognition from Parallel Corpora

Andreea Bodnari[1,2], Aurélie Névéol[2], Ozlem Uzuner[1,3], Pierre Zweigenbaum[2], and Peter Szolovits[1]

[1] MIT, CSAIL, Cambridge, Massachusetts, USA
[2] LIMSI-CNRS, rue John von Neumann, F-91400 Orsay, France
[3] Department of Information Studies, University at Albany, SUNY, Albany, New York, USA

Abstract. We present a named-entity recognition (NER) system for parallel multilingual text. Our system handles three languages (i.e., English, French, and Spanish) and is tailored to the biomedical domain. For each language, we design a supervised knowledge-based CRF model with rich biomedical and general domain information. We use the sentence alignment of the parallel corpora, the word alignment generated by the GIZA++[8] tool, and Wikipedia-based word alignment in order to transfer system predictions made by individual language models to the remaining parallel languages. We re-train each individual language system using the transferred predictions and generate a final enriched NER model for each language. The enriched system performs better than the initial system based on the predictions transferred from the other language systems. Each language model benefits from the external knowledge extracted from biomedical and general domain resources.

Keywords: Named-entity recognition, Medical NLP

1 Introduction

Natural Language Processing (NLP) technologies can help extract structured information from written text, determine relationships between concepts, or perform bilingual translation. In the biomedical and clinical[4] domain NLP has multiple applications: computerized clinical decision support, personalized medicine, automatic extraction of protein interactions and associations of proteins to functional concepts, to name a few. Despite their value and extensive applicability, biomedical and clinical NLP systems are scarcely developed for languages other than English.

In this study, we present a system for extracting structured information from multilingual biomedical corpora. We define the structured information of interest as noun phrases (NPs) that fall under nine Unified Medical Language System (UMLS)[2] semantic categories. Our data consists of parallel multilingual corpora in English (en), French (fr), and Spanish (es).

We present a NP extraction system for multilingual biomedical corpora that makes use of external knowledge sources and the common structure shared by the parallel languages. We analyze the errors performed by the NER system, and we discuss the contribution of the knowledge sources and the common language structure to system performance.

1.1 Problem definition

We focus on NPs that fall under nine semantic categories: anatomy (ANAT), chemicals and drugs (CHEM), devices (DEVI), disorders (DISO), geographic areas (GEOG), living beings (LIVB), objects (OBJC), phenomena (PHEN), and physiology (PHYS). These semantic categories are groupings of semantic types in the UMLS semantic network.[7] Our goal is two-fold: for each of the three languages, first identify noun phrases that belong to the semantic categories of interest and then map the identified noun phrases to an already existing Concept Unique Identifier (CUI) in UMLS.

2 Materials and Methods

2.1 Data

The data used for the system development is made available by the organizers of the 2013 CLEF-ER challenge[9] and comes from two sources: the European Medicines Agency (EMEA) and Medline. Each corpus contains sentence-delimited plain text. Sentence alignment information is available for the language pairs en-fr and en-es. The EMEA corpus contains the same number of sentences for each language (140522) so that each English sentence has an equivalent alignment in the other two languages. The Medline corpus contains approx. 1.5 million English sentences, 0.5 million French sentences and 0.25 million Spanish sentences (see Table 1). In the Medline corpus, half of the English sentences do not have an equivalent alignment in French or Spanish, while the other half have an equivalent alignment in either French or Spanish.

	Sentence count		
	English	French	Spanish
EMEA	140552	140552	140552
Medline	1593546	572176	247655

Table 1. Corpus sentence count per source type, separated by language

2.2 Annotations

We manually annotated all noun phrase instances of the nine semantic types within a set of 385 sentences, for each language and for each corpus source. We use different annotation processes for the EMEA and the Medline corpora. For the EMEA corpus, we randomly selected 385 sentences from the automatically generated gold standard provided by the challenge organizers. For each of the selected sentences, we transferred the English annotations to the corresponding French- and Spanish-aligned sentences based on the word alignment provided by the GIZA++ software and Wikipedia as explained below. We manually reviewed the transferred annotations and annotated additional noun phrases for each language when appropriate.

For the Medline corpus, we used the Medical Subject Heading (MeSH) indexing and the UMLS Metathesaurus to automatically generate noun phrase annotations. The MEDLINE citations corresponding to the sentences (i.e., titles) in the Medline corpus have been manually assigned a set of MeSH indexing terms by professional indexers at the National Library of Medicine. MeSH main headings can be derived from this set and mapped to UMLS CUIs. We randomly selected 385 Medline sentences for each language and made use of the MeSH indexing-UMLS CUI relationship in order to generate Medline reference annotations. Specifically, when a term associated to a CUI from the MeSH indexing can be found in the corresponding Medline sentence, we automatically create a reference annotation. We apply the same annotation process to the English, French, and Spanish versions of the Medline corpus.

The EMEA and Medline reference annotations were manually reviewed by a sole annotator, due to time constraints and lack of additional resources.

The number of unique annotations for EMEA and Medline are similar across languages (368 English annotations for EMEA and 401 for Medline, 364 French annotations for EMEA and 339 for Medline, and 368 Spanish annotations for EMEA and 303 for Medline). The EMEA corpus contains a significantly larger number of actual annotations due to numerous noun phrase duplicates (see Table 2).

2.3 System design

We perform word alignment using the GIZA++ software.[8] GIZA++ aligns words based on statistical models. For source sentence $s^J = s_1 s_2 .. s_J$ and target sentence $t^I = t_1 t_2 .. t_I$, we define an alignment of the two sentences as $A \subseteq \{(i,j) : j = 1,..,J; i = 0,..,I\}$, where the case $i = 0$ represents source words that are not aligned to any target words. The probability of a source sentence given a target sentence is $P(s|t) = \sum_{a_J} P(s_J, a_1^J | t^I)$, where a_1^J represents the sentence pair alignment. The best sentence alignment is given by $\hat{a}^J = argmax_{a^J} P_\theta(s^J, a^J | t^I)$.

In addition to the GIZA++ alignment, we use word alignment from Wikipedia metadata. We rely on the fact that some English Wikipedia articles have direct correspondents in French and Spanish. We filter the English articles that have titles consisting of a single word (e.g., “Food”, “Disease”, “Nausea”) and find their corresponding foreign language article. The title of the English article and the title of the correspondent foreign language article represent a word alignment pair. Because Wikipedia word alignments are more precise, we overwrite the GIZA++ generated word alignments with a score lower than 0.5 with the Wikipedia generated word alignments.

		Annotation count		
		English	French	Spanish
ANAT	EMEA	202	31	28
	Medline	62	60	38
CHEM	EMEA	1556	519	572
	Medline	88	55	58
DEVI	EMEA	43	6	10
	Medline	10	3	1
DISO	EMEA	1658	427	465
	Medline	168	180	174
GEOG	EMEA	34	11	13
	Medline	15	9	22
LIVB	EMEA	492	132	137
	Medline	72	60	35
OBJC	EMEA	115	37	38
	Medline	9	4	2
PHEN	EMEA	77	30	30
	Medline	2	2	1
PHYS	EMEA	272	53	64
	Medline	29	12	6
Total NPs	EMEA	4491	1261	1374
	Medline	461	386	340
Total Unique NPs	EMEA	368	364	368
	Medline	401	339	303

Table 2. Annotations description per corpus type: annotation count, separated by language and semantic category

Noun phrase identification Our system is designed based on the supervised CRF framework. The CRF model includes lexical features (the normalized form of the token), syntactic features (part-of-speech and parse tree information of each token), and knowledge-based features (information extracted from UMLS and Wikipedia). A final set of features is generated based on the shared structure of the parallel languages. All feature sets are used in the final CRF model.

In order to obtain the parse and POS information we use different resources for each language: for English we use the Stanford parser,[5] while for French and Spanish we use the Malt parser.[3, 6] We extract the knowledge-based features from UMLS using MetaMap[1] for the English corpus and direct lexical mapping for the French and Spanish corpora. The lexicons we used contained about 0.3 million terms corresponding to 0.15 million CUIs for French and 0.5 million terms corresponding to 0.3 million CUIs for Spanish.

The Wikipedia knowledge-based features are dependent on the language and are extracted based on the respective language version of Wikipedia. We make use of the fact that Wikipedia articles are tagged with specific categories. We map these Wikipedia categories to one of the nine UMLS categories of interest. Then, we identify the n-grams ($n \leq 3$) that represent titles of Wikipedia articles and signal within the attribute vector whether an n-gram belongs to one of the nine categories of interest based on its Wikipedia categorization.

Using the lexical, syntactic, and knowledge-based features described above, we create a CRF model (the Initial model) for each language and label the training data using the relevant model. We further make use of the common structure shared by the parallel corpora and transfer the NP predictions across aligned bilingual sentence pairs. We hypothesize that some language models might predict NPs that other language models would miss, and by using sentence- and word-alignment information we can inform the other language models of the missed predictions. The CRF model containing the transferred predictions is referred to as the enriched model. The entire system design is depicted in Figure 1.

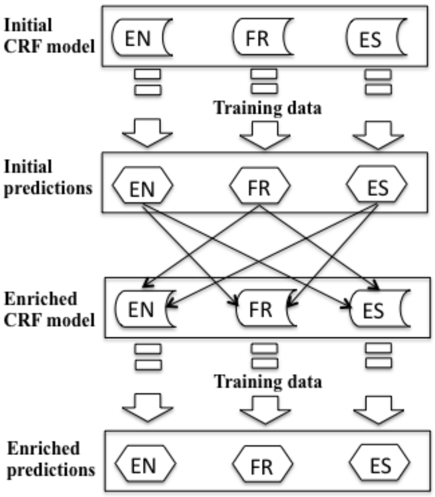

Fig. 1. Multilingual NP extraction: system description

2.4 CUI Mapping

The NER system predicts the text of the NPs together with their semantic categories. In order to obtain the CUI associated with each NP we use MetaMap for English and direct lookup inside the UMLS database for the French and Spanish languages. Because the French and Spanish versions of UMLS do not contain the same number of concepts as the English UMLS, there are French and Spanish concepts identified by our system that cannot be linked to a CUI via direct UMLS lookup. For those concepts we rely on word alignment to determine the English concept to which the foreign language concepts are aligned; once an aligned English concept is found we transfer the CUI information to the foreign language concept.

3 Discussion

The reference annotations we generated for this challenge relied partially on automated annotation tools. These annotation tools generated an imperfect annotation output and generally failed to identify a percentage of the noun-phrase instances. For the Medline corpus, the annotation tools failed to identify NPs in the sentences that belonged to oldest citations. These citations were only assigned a couple of MeSH indexing terms, vs. about a dozen for more recent citations. Also, because the MeSH indexing terms are assigned based on the full-text article content and not on the article title, they might be linked to CUIs not present in the title. A similar issue arises when the Medline title contains a shortened or altered version of the CUI string indexed by MeSH (e.g. the string "hypertensive patient" occurring in the title of citation 1838917 could not be reconciled with CUI C0020538 "hypertension"). For EMEA, the main problem was the transfer of English annotations based on the word alignment: the GIZA++ software does not generate a perfect word alignment, and even though the Wikipedia word alignment manages to fix some incorrect alignments, there are still cases of incorrect or missing alignments.

The problem of the noisy word alignment impacts both the generation of the reference annotations and the performance of the enriched system. Specifically, the predictions transferred between the systems are usually incomplete (e.g., for the English noun phrase active substance, the word alignment could only help transfer the labeling for the token "substance" and failed to transfer the labeling for the token "active" as it could not find a valid word French or Spanish alignment for the word "active").

Mapping the noun phrases to a CUI is a difficult problem for the French and Spanish languages since the number of annotated CUIs is relatively small. A large number of UMLS English concepts are not present in the foreign language portions of UMLS; thus, even though the English translation of a French or Spanish noun phrase is present in UMLS, the French or Spanish noun phrase is not included. Our system has to rely on machine translation or noun phrase alignment for mapping certain French and Spanish noun phrases to their CUIs.

Based on manual revision, the noun phrases identified by the enriched model present a higher precision and slightly lower recall. In general, the enriched model predicts more complete noun phrases (e.g., "insuffisance hépatique sévère" vs. "insuffisance hépatique", "diminution des bruits respiratoires", vs. "bruits respiratoires", "parésies des cordes vocales" vs. "des parésies des cordes", "tumor principal" vs. "tumor", "movimientos involuntarios" vs. "involuntarios"). We notice that the semantic categories of the noun phrases are better assigned when the enriched model is used (e.g., "linfocitos" assigned a LIVB category by the initial model and an ANAT category by the enriched model), and even correctly adjusted together with the span of the noun phrase (e.g., noun phrase "mandíbula" classified as ANAT in the initial model is changed into osteonecrosis de la mandbula classified as DISO in the enriched model).

4 Conclusion

We present a multilingual NER extraction system targeting the biomedical domain. The NER system relies on a supervised learning framework enriched with external knowledge and cross-linguistic information transferred based on common structure between languages. We prepare reference annotations for English, French, and Spanish that together with the NER system can be used for developing additional multilingual resources. We illustrated the improvements brought by our second round of training based on cross-language transfer of initial annotations.

5 Acknowledgments

This work was funded in part by award number 2U54LM008748 from the National Institutes of Health (NIH)/National Library of Medicine (NLM) and cofounded by the National Heart, Lung and Blood Institute (NHLBI), and by contract number 90TR0002 (SHARPd Secondary Use of Clinical Data) from the Office of the National Coordinator (ONC) for Health Information Technology. The main author was funded by the Chateaubriand Science Fellowship 2012-13.

References

1. Alan R Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.
2. Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.
3. Marie Candito, Benoît Crabbé, Pascal Denis, et al. Statistical french dependency parsing: treebank conversion and first results. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1840–1847, 2010.
4. Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772, 2009.
5. Dan Klein and Christopher D Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics, 2003.
6. Montserrat Marimon, Núria Bel, Sergio Espeja, and Natalia Seghezzi. The spanish resource grammar: pre-processing strategy and lexical acquisition. In *Proceedings of the Workshop on Deep Linguistic Processing*, pages 105–111. Association for Computational Linguistics, 2007.
7. Alexa T McCray, Anita Burgun, Olivier Bodenreider, et al. Aggregating UMLS semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, (1):216–220, 2001.
8. FJ Och and H Ney. Giza++: Training of statistical translation models, 2000.
9. Hanna Suominen, Sanna Salantera Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In *Proceedings of CLEF 2013*, 2013.