Machine Translation for Entity Recognition across Languages in Biomedical Documents

Giuseppe Attardi, Andrea Buzzelli, Daniele Sartiano

Dipartimento di Informatica
Università di Pisa
Italy
{attardi, buzzelli, sartiano}@di.unipi.it

Abstract. We report on our experiments for the CLEF 2013 Entity Recognition Challenge. Our approach is based on a combination of machine translation and NE tagging techniques. The Silver Standard Corpus (SSC) is used to obtain a corresponding annotated corpus in the target language. The plain text of the SSC is translated and a mapping is created between entities in the original and phrases in the translation, to which are associated the same CUIs as in the original. This produces a Bronze Standard Corpus (BSC) in the target language. A dictionary of entities is also created, which associates to each pair (entity text, semantic group) the corresponding CUIs that appeared in the SSC. The BSC is used to train a model for a Named Entity tagger. The model is used for tagging entities in sentences in the target language with the proper semantic group and the entity dictionary is used for associating CUIs to each of them.

1 Introduction

The CLEF-ER Challenge 2013 targets the task of multilingual identification of mentions of named entities in several corpora. Patent texts, titles of Medline abstracts and EMEA documents serve as corpora.

The task is to produce mention annotations in a multilingual document, i.e. assignment of Concept Unique Identifiers (CUIs, from the UMLS) to phrases in the corpus.

The challenge organizers provide several resources such as

1. a terminological resource (TR) produced from UMLS, containing English and non-English concepts together with their CUIs.
2. a selection of corpora in English, i.e. patent texts, Medline titles and EMEA documents, where the entity mentions have been annotated automatically with their CUIs.
3. a selection of corpora in different languages other than English, i.e. in DE, FR, SP, and NL, that have to be annotated with entity mentions and their CUIs.

The English corpora have been annotated automatically with the help of several annotation solutions (for English) from the project partners and a harmonisation scheme to

generate a Silver Standard Corpus (CALBC approach). The English corpus thus serves as additional input, but does not serve as a Gold Standard.

2 Approach

We designed our system while we did not have yet access to the UMLS TR resources provided by the organizers, therefore we just exploited the annotated English Silver Standard Corpus (SSC) as a source of information. We will discuss later how the TR resources could be integrated in our approach.

Our approach combines techniques of machine translation with NER techniques.

The following is an overview of the approach:

1. we apply phrase-based statistical machine translation to the SSC in order to obtain a corresponding annotated corpus in the target language. The plain text of the SSC is translated and a mapping is created between entities in the original and phrases in the translation, to which are associated the same CUIs as in the original. This produces a Bronze Standard Corpus (BSC) in the target language. A dictionary of entities is also created, which associates to each pair (entity text, semantic group) all the corresponding CUIs that appeared in the SSC.
2. the BSC is used to train a model for a Named Entity tagger, whose output classes are the possible semantic groups of entities.
3. the model built at step 2) is used for tagging entities in sentences in the target language with the proper semantic group.
4. the annotated document is converted to XML format and enriched by adding CUIs to each entity, looking up the pair (entity, group) in the dictionary of CUIs built in step 1.

One advantage of this approach is that it produces data, a NER model and an entity dictionary, that can be readily applied to a document in the target language without any further reference to the source corpora in the original language.

2.1 Creating the Bronze Standard Corpus

The Bronze Standard Corpus is obtained by translating the SSC and transferring to it the entity annotations.

For translating the original English SSC into the target language (Spanish), we use Moses [7], a statistical phrase-based machine translation system that allows automatically learning translation models for any language pair. Moses is trained through a collection of parallel corpora. An efficient decoder algorithm finds the highest probability translation among an exponential number of possible translations.

We exploited the word alignment information produced by Moses to determine the correspondence between entities in the source and target sentences.

Moses was trained on texts pertaining to the biomedical domain, obtained by joining the EMEA corpus [5] with from the Medline resource provided for the CLEF-ER challenge.

In our experiments we only dealt with English to Spanish translation, but the approach can be applied to any language pair for which there exist suitable parallel corpora.

In order to identify in the target language the phrases that correspond to entities in the original, we exploit the word alignment information obtained by invoking the Moses decoder with the option:

-alignment-output-file file

Let's illustrate this step with the following example.

"This medicine/CHEM relieves headaches/DISO"

By invoking the Moses decoder, we obtain the following best translation:

"Este medicamento alivia los dolores de cabeza"

and the following word alignment:

1-1 2-2 3-3 3-4 3-5 3-6

Each pair of numbers in the alignment provides a correspondence between a token in the original sentence identified by its left number position and a token in its translation identified by its right number position.

The word alignment allows us to map an entity in the source to its translation.

For example, the entity "medicine", located at position '1' in the original sentence, is mapped to the single word "medicamento", at position '1' in the translation.

The case is less clear for the second entity: the word alignment indicates that the source word "headaches", located at position '3', maps to the list of tokens [3, 4, 5, 6], leading to the phrase "los dolores de cabeza", as a possible entity text in the target language. This candidate text is cleaned up by removing articles and punctuations that occur at the beginning or end of the entity text, in order to obtain more consistent phrases.

In the example, article "los" gets dropped from the beginning, producing the final entity "dolor de cabeza". This step is performed by simply checking the part of speech tags of tokens, obtained by using the Tanl POS tagger [1], trained for Spanish on the Ancora corpus [4].

2.2 Building the NER Training Set

The training set for the NE tagger is obtained from the BSC, converting it into IOB notation and adding POS tags to each token.

The example in previous section would be represented as follows:

FORM	POS	IOB
Este	DD	O
medicamento	NC	B-CHEM
alivia	VM	O

los	DA	O
dolores	NC	B-DISO
de	SP	I-DISO
cabeza	NC	I-DISO

2.3 Training the NE tagger

For performing NE recognition, we used the *Tanl Tagger* [1], a generic, customizable statistical sequence labeller, suitable for many tasks of sequence labelling, such as POS tagging or Named Entity recognition. The tagger implements a Conditional Markov Model (CMM, aka MEMM) [2] for sequence labeling that combines features of Hidden Markov models (HMMs) and Maximum Entropy models.

The Tagger can be configured to use alternative types of classifiers: Maximum Entropy or Linear Support Vector [6]. By complementing the classifier with dynamic programming the Tanl Tagger can achieve similar levels of accuracy than SVM with much faster speed.

2.4 Mention Identification

In order to identify mentions of named entities in a document in the target language, we can apply the previously built NE tagger model and entity dictionary.

The input for the NER tagger must be prepared performing sentence splitting (if needed), tokenization and POS tagging, producing a TSV file for the NE tagger, containing two fields for each token, FORM and POS tag.

The NER tagger applies the previously build model, producing a three column file, with the third column denoting with the semantic group, in the IOB notation.

Finally CUIs are added to each mention by looking it up in the previously built dictionary of entities and CUIs. The output is then formatted in the *standoff* notation required for the task..

Notice that the process could actually be performed fully in memory, using the Tanl pipeline, passing tokens from one stage to the next, without producing any intermediate file.

3 Experiments

3.1 Data

For training moses we used a parallel corpus consisting of 247,655 sentences from the English-Spanish version of Medline and 1,098,333 sentences from the EMEA corpus [5]. We tokenized the corpus with the Tanl Tokenizer [1] and then split into a train part with 1,323,588 sentences and a development part with 11,000 sentences.

3.2 Moses training

We ran the training of Moses, using the *KenLM* [9] language model, created using the text of the text of the whole Spanish Wikipedia, extracted using the Tanl Wikipedia Extractor tool [8].

Word alignment was done using MGIZA, a multi thread version of GIZA++ [10]. After training, we performed a tuning process using the development corpus, created as mentioned earlier.

The Moses decoder was run using this model with default settings except for a beam size of 500.

3.3 NER Training

The Tanl NE tagger can be customized by specifying the set of features to use. Features are divided into local and global features. Local features include *attribute features*, extracted from attributes (e.g. Form, PoS, NE) of tokens in the vicinity of current token, and *morphological features*, binary features extracted from a token matching a given regular expression. For CLEFER we did not use any *global feature*: properties holding at the document level. We tested several configurations, starting from the one that proved best for Italian at Evalita 2011 [3]. The submitted official run uses the same morphological features and these attribute features:

Attribute	Token Position
FORM	-2 -1 0
POSTAG	-1 0 1
NETAG	-1

For CLEFER we used a Linear SV classifier.

4 Results

We submitted three runs for evaluation. The official submission (EMEA_es_man.LR-5.xml) identified 417,390 entities in the 140,552 units present in the Spanish test corpus EMEA_es_man.xml.

Some of the identified entities appear without the corresponding CUI while others appear with a very large number of CUIs. The former occurs typically for entities with a general meaning, that during translation get associated to several different original English words in the entity dictionary.

After the submission, we were able to address this problem by performing a clean-up of the entity dictionary, in this way:

```
for each pair (e, cl) in the dictionary
   for each c in cl
      retrieve the set of text entities te associated
         to c in the UMLS for target language
```

`if e is not present in te remove it from cl`

5 Conclusions

We reported our experiments in the CLEF-ER Challenge 2013 for English to Spanish. By exploiting phrase-base machine translation and NE tagging we were able to build a system that can operate standalone, on any text in the target language, effectively transferring the knowledge on entities from one language to another.

The accuracy of the solution can be further refined by making use of data from the UMLS, as we started doing in our latest experiments.

Acknowledgements. Partial support for this work was provided by project RIS (POR RIS of the Regione Toscana, CUP n° 6408.30122011.026000160).

References

1. Attardi, G., Dei Rossi, S., Simi, M.: The Tanl Pipeline. In: *Proc. of Workshop on Web Services and Processing Pipelines in HLT*, Malta (2010)
2. McCallum, A., Freitag, D., Pereira, F.: Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proc. ICML 2000*, 591–598 (2001)
3. G. Attardi, G. Berardi, S. Dei Rossi, M. Simi. The Tanl Tagger for Named Entity Recognition on Transcribed Broadcast News at Evalita 2011. In B. Magnini et al. (Eds.), *Proc. of Evalita 2011*, LNCS 7689, pp. 116-125, 2013. ISBN 978-3-642-35827-2.
4. M. Civit Torruella and M. A. Martì Antonìn. 2002. Design Principles for a Spanish Treebank, In *Proc. of the First Workshop on Treebanks and Linguistic Theories* (TLT).
5. Tiedemann, J.: News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R., eds.: *Recent Advances in Natural Language Processing*. Volume V. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria (2009) 237–248
6. Yu, H. F., Hsieh, C. J., Chang, K. W., Lin, C. J.: Large linear classification when data cannot fit in memory. ACM Trans. on Knowledge Discovery from Data, 5:23:1-23 (2012)
7. Koehn, Philipp, et al. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. ACL, 2007.
8. Tanl Wikipedia Extractor. http://medialab.di.unipi.it/wiki/Wikipedia_Extractor
9. Heafield, Kenneth. KenLM: Faster and smaller language model queries. In *Proc. of the Sixth Workshop on Statistical Machine Translation.* ACL, 2011.
10. Och, Franz Josef, and Hermann Ney. Giza++: Training of statistical translation models. (2000).