Automatic Generation of Named Entity Recognition Taggers Using Parallel Corpora[*]

Generación de Etiquetadores de Entidades Nombradas mediante Alineamientos de Palabras

Rodrigo Agerri, Itziar Aldabe, Nora Aranberri
Yiling Chung, Gorka Labaka, German Rigau
IXA NLP Group
Euskal Herriko Unibertsitatea UPV/EHU
rodrigo.agerri@ehu.eus

Resumen: The lack of hand curated data is a major impediment to developing statistical semantic processors for many of the world languages. Our paper aims to bridge this gap by leveraging existing annotations and semantic processors from multiple source languages by projecting their annotations via statistical word alignments traditionally used in Machine Translation.
Palabras clave: Semántica, Alineamientos, Extracción de Información.

Abstract: La ausencia de datos de entrenamiento manualmente anotados es un impedimento fundamental para el desarrollo de procesadores semánticos estadísticos para la mayoría de los idiomas. Este artículo tiene como objetivo generar automáticamente procesadores semánticos multilingües mediante mediante alineamientos de palabras (tradicionalmente usadas en Traducción Automática) a un idioma para el cual no se dispone de datos manualmente anotados.
Keywords: Semantics, Alignments, Information Extraction.

1 Introduction

A major issue of Named Entity Recognition (NER) taggers is that they require manually annotated data to perform accurately. In this work we propose a method to automatically induce Named Entity taggers using parallel data, without any manual intervention.

Our method leverages existing semantic processors and annotations to overcome the lack of annotation data for a given language. The intuition is to transfer or project semantic annotations, from multiple sources to a target language, by statistical word alignment methods applied to parallel texts (Och and Ney, 2000; Liang, Taskar, and Klein, 2006). The projected annotations could then be used to automatically generate semantic processors for the target language. In this way we would be able to provide NLP processors without training data for the target language.

Thus, this means that the problem can be decomposed into two smaller inter-related ones: (i) How to project semantic annotations across languages via parallel texts with a sufficient acceptable quality to train semi- or weakly-supervised semantic processors and (ii) how to effectively leverage the (potentially noisy) projected annotations to induce robust statistical models to perform semantic tasks such as NER, WSD or SRL.

In this paper we focus on the first problem. We propose using parallel data from multiple languages as source to project the semantic annotations to a target language. Our hypothesis is that in the combination of multiple sources lies the possibility of improving the quality of the projections that will be used to train the semantic processors.

2 Experiments

The experiments are focused on 4 languages: German, English, Spanish and Italian. For the first three languages we use the well-known CoNLL 2002 and 2003 datasets (Tjong Kim Sang, 2002). For Italian, we use the Evalita 2009 dataset. Both CoNLL and

[*] This work has been funded by the Spanish Ministry for Science and Innovation (MICINN) TUNER, TIN2015-65308-C5-1-R. and CROSSTEXT, TIN2015-72646-EXP

Training Data	English	German	Italian	Spanish
Gold	65.08	49.87	65.82	58.75
Projected	69.14	70.62	62.44	64.16

Table 1: Evaluating projected and gold-standard models on Europarl test.

Evalita annotate the three entity types (location, organization and person) that we will use to induce our training data. For parallel data we chose Europarl (Koehn, 2005), which was word-aligned using Giza++ (Och and Ney, 2000) and divided in a training and test set. The test set contains 800 sentences manually annotated using the three entity types and following the CoNLL 2002 and 2003 guidelines. Our method consists of the following four steps:

1. We train ixa-pipe-nerc (Agerri and Rigau, 2016) on the gold-standard training data from CoNLL and Evalita.
2. The Europarl training data for each language is tagged with the gold-standard trained models.
3. We project the automatic tagged named entities from three source languages to a fourth target language.
4. ixa-pipe-nerc is then trained on the induced training data via projection across languages obtaining a NER tagger which is fully automatically generated.

The projection of the named entity annotations using parallel data uses both the automatically obtained named entity and word alignments from the Europarl training set. Firstly, given a word in a sentence of the target language, we obtain the aligned words and their named entity class in the three source languages. Next, the named entity tags of target language are projected based on the candidates collected from the three source languages. For the first version of our projection system we aim at high precision, so we developed a *strict-match* projection algorithm that considers at least two or three alignment agreements among three source languages to determine the final tag for target language. If that agreement is not reached, we use a back-off named entity tag obtained from computing the most frequent tag for that token in Wikiner, a large automatically annotated corpora (Nothman et al., 2013).

As we have already mentioned, we compare the gold-trained models with the automatically induced ones on our Europarl gold-standard. This evaluation allows to understand if our method produces as good results as the models trained on gold standard, albeit out-of-domain, data. The F1 results in Table 1 show that the automatically trained models outperform the models trained on gold-standard data except for Italian. Furthermore, our automatically obtained models are particularly good in terms of precision, which means that our strict match projection algorithm is very strict, and only projects named entities when it is quite sure. Thus, for English the precision results are 6 points higher, 22 points for Spanish and 6 points for German.

3 Concluding Remarks

We train the same tagger on the automatically created training data and on out-of-domain gold-standard annotated data. Our evaluation shows that the automatic generated model outperforms the gold-standard trained model in an in-domain evaluation. Our work shows how to automatically induce training data using parallel data without manual intervention. This method allows to generate named entity taggers for a given language when no manually data is available. Furthermore, our method can be applied to generate annotations for other semantic tasks, such as Semantic Role Labeling or Supersense tagging.

Future work includes evaluating both gold-trained and projected models on out-of-domain data using the MEANTIME corpus (Minard et al., 2016). After all, NER taggers are usually used to tag out-of-domain data, so if our automatically generated models were to be at least as good as the models trained on gold-standard out-of-domain data, that would mean that for out-of-domain use our method would be a convenient solution to obtain general semantic processors without manual intervention.

Bibliografía

Agerri, R. and G. Rigau. 2016. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63–82.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5.

Liang, P., B. Taskar, and D. Klein. 2006. Alignment by agreement. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 104–111.

Minard, A.-L., M. Speranza, R. Urizar, B. Altuna, M. van Erp, A. Schoen, and C. van Son. 2016. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In *Proceedings of LREC 2016*.

Nothman, J., N. Ringland, W. Radford, T. Murphy, and J. R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Och, F. J. and H. Ney. 2000. *Giza++: Training of statistical translation models*.

Tjong Kim Sang, E. F. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.