Building Named Entity Recognition Taggers via Parallel Corpora

Rodrigo Agerri, Yiling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka, German Rigau
IXA NLP Group, University of the Basque Country UPV/EHU
rodrigo.agerri@ehu.eus,lydia193@gmail.com,{itziar.aldabe,nora.aranberri,gorka.labaka,german.rigau@ehu.eus}

Abstract

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 the statistical word alignments traditionally used in Machine Translation. Taking the Named Entity Recognition (NER) task as a use case, this work presents a method to automatically induce Named Entity annotated data using parallel corpora without any manual intervention. The projected annotations can then be used to automatically generate semantic processors for the target language helping to overcome the lack of training data for a given language. The experiments are focused on 4 languages: German, English, Spanish and Italian, and our empirical evaluation results show that our method obtains competitive results when compared with models trained on gold-standard, albeit out-of-domain, data. The results point out that our projection algorithm is effective to transport NER annotations across languages thus providing a fully automatic method to obtain NER taggers for as many as the number of languages aligned in parallel corpora. Every resource generated (training data, manually annotated test set and NER models) is made publicly available for its use and to facilitate reproducibility of results.

Keywords: Named Entity Recognition, Information Extraction, Multilingual Language Resources

1. Introduction

The best results for every type of semantic processing task are currently obtained by supervised corpus-based approaches. This means that manually annotated data is required to learn probabilistic models from the data. This poses a major obstacle to developing semantic processors whenever there is not manually annotated data for a semantic task in a given language. In most cases, manually annotating text for every single specific need is generally inefficiently slow and, in most cases, unaffordable in terms of human resources and economic costs. Instead, we would like to be able to use already available semantic processors and texts in other languages to get a good statistical model for a new target language.

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 et al., 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 semantic processors without training data for a given 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 Named Entity Recognition (NER), Word Sense Disambiguation or Semantic Role Labelling, to name but a few.

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. For the purpose of this work, we take the NER task as a use case to test our hypothesis. Furthermore, four languages are considered in our study: English, Spanish, German, and Italian, although any language aligned in a a parallel corpus is a possible candidate. Our method can be illustrated by the following example provided in Figure 1, which takes English as a target language.

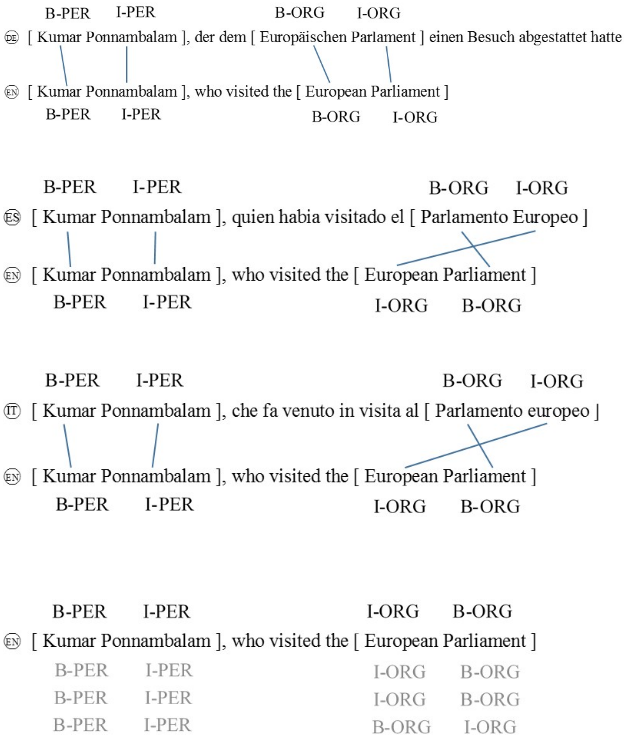
B-PER I-PER B-ORG I-ORG
DE [Kumar Ponnambalam], der dem [Europäischen Parlament] einen Besuch abgestattet hatte
EN [Kumar Ponnambalam], who visited the [European Parliament]
B-PER I-PER B-ORG I-ORG
B-PER I-PER B-ORG I-ORG
ES [Kumar Ponnambalam], quien habia visitado el [Parlamento Europeo]
EN [Kumar Ponnambalam], who visited the [European Parliament]
B-PER I-PER I-ORG B-ORG
B-PER I-PER B-ORG I-ORG
IT [Kumar Ponnambalam], che fa venuto in visita al [Parlamento europeo]
EN [Kumar Ponnambalam], who visited the [European Parliament]
B-PER I-PER I-ORG B-ORG
B-PER I-PER I-ORG B-ORG
EN [Kumar Ponnambalam], who visited the [European Parliament]
B-PER I-PER I-ORG B-ORG
B-PER I-PER I-ORG B-ORG
B-PER I-PER B-ORG I-ORG

Figure 1: Projecting Named Entity annotations via word alignments to English as target language.

Alignments	Tokens in source languages es; de; it	Target language (en)
1-1 alignment	Europeos; Europas; europeo	European
Multiple alignments	del, Parlamento; Parlamentsgebäude; Parlamento	Parliament
Misalignments	del; Parlamentsgebäude; Parlamento	Parliament
No alignments	Los; Beschäftigungspakten; NONE	European

Table 1: Examples of various alignments from Spanish, German, Italian to English.

2. Methodology

In order to develop our system we need: (i) a Named Entity Recognition (NER) tagger; (ii) a parallel corpus to project the semantic annotations in order to create the training data; (iii) NER datasets for training the initial models to be deployed to tag the parallel corpus, and (iv) a gold-standard test set to evaluate our approach.

As NER tagger we choose *ixa-pipe-nerc* (Agerri and Rigau, 2016). It is designed to work robustly across languages and datasets and it obtains state of the art results for the languages used in this study. We also use the following corpora:

1. Gold standard data for training the initial *ixa-pipe-nerc* models for the source languages. CoNLL 2002 and 2003 for German, English and Spanish, and Evalita 2009 for Italian. Both CoNLL and Evalita annotate the three entity types (location, organization and person) that we will use to induce our training data.

2. The Europarl parallel corpus on which to perform the cross-lingual projections (Koehn, 2005), word-aligned using Giza++ (Och and Ney, 2000) and divided into a training and a test set.

3. The Europarl gold-standard test set is a new manually-annotated evaluation set taken from the Europarl. The test set contains 800 sentences manually annotated using the three entity types and following the CoNLL 2002 and 2003 guidelines for the 4 languages used in this paper.

4. Back-off corpora to resolve ties in the projection step. The idea is to compute the most frequent tag of a token in a large NER annotated resource. Thus, in case of ties during the annotation projection the most frequent entity tag will be assigned. We use Wikiner, a silver standard built from the Wikipedia for several languages (Nothman et al., 2013).

3. Projecting Annotations

There can be various types of word alignments between the tokens and Named Entity (NE) classes for any two languages: 1-1 alignments, multiple alignments, misalignments and no alignments. Examples of such cases can be found in Table 1:

- 1-1 alignments occur when the NE class of a token in the source is cleanly aligned to one token and NE class in the target.

- Multiple alignments means that two or more tokens and NE classes are aligned from the source to one token and NE class in the target (e.g., *del* and *Parlamento* in Spanish aligned to one token *Parliament* in English).

- Misalignments happen when a token is wrongly aligned to a token in the target. Our algorithm considers different strategies for these different types of (mis-)alignments.

- No alignments occur when the token containing a NE class in the source is not aligned to any token in the target.

Taking this into account, for this work we have developed two projection algorithms: (1) an *upper bound* designed to evaluate the quality of the word alignments to transport gold-standard Named Entity Recognition (NER) annotations; (2) *strict match* projection algorithm to project automatically annotated NER annotations over the training set. The automatically projected annotations will then be used to train new NER models for the target language.

3.1. Upper-Bound

In order to establish the quality of the word alignments to project semantic annotation, we designed an *upper bound* projection method. In order to do so, we project the manually annotated Named Entities using the Europarl test set described in Section 2. Furthermore, we only transport a NE tag to the target language whenever all three source annotations coincide. Thus, no back-off will be used for the upper-bound. The resulting projected data will be evaluated with respect to the test gold-standard of the target language using the CoNLL script for NER evaluation. Intuitively, the results should establish how much noise is created by the projection via word alignments whenever the alignments and tags for all the three source languages agree. For each language, the upper-bound project algorithm performs the following steps:

1. We obtain the word alignments for all four languages and order them following the alignment types presented in Table 1.

2. The semantic tags are projected via word alignments from the three source languages to a given fourth, target language. In this step a NE tag is projected only if there are three agreements between the alignments in the source. This may happen whenever there is a 1-1 alignment, or if there are multiple alignments but three of the tags coincide across the three source languages. Otherwise, the projected tag is ‘O’. This is illustrated by Table 2.

Alignments	Tokens	Tags in projection es; de; it	Projected tag
1-1 alignment	European	ORG; ORG; ORG	ORG
Multiple alignments	Parliament	ORG; O, ORG; ORG	ORG

Table 2: Examples of upper-bound projection for English as target language.

3. Assign the span to the projected NE tag: It should be noted that the projections are performed at token level, so in order to annotate a Named Entity in the target language, the projected NE tags must be contiguous.

3.2. Strict Match

The aim of our work is to project Named Entity (NE) annotations from several source languages into a target language for which there is not training data. Thus, the result of the projection will be an automatically created training corpus on which we could train a Named Entity Recognition (NER) tagger. As there are not large parallel corpora manually annotated with Named Entities on which to perform the projections, we annotate a large parallel corpus with already existing Named Entity taggers so that we can use the automatic Named Entity annotations and word alignments to undertake the projection across languages. Specifically, in order to automatically generate NER taggers without manual intervention via the *strict-match* algorithm our method goes through the following four steps:

1. We train ixa-pipe-nerc (Agerri and Rigau, 2016) on the gold-standard training data from CoNLL (Tjong Kim Sang and De Meulder, 2003; Tjong Kim Sang, 2002) and Evalita (Speranza, 2009).

2. The Europarl training set 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, performing all 4 permutations.

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.

For *strict match*, given a word in a sentence of the target language, we obtain the aligned words and its NE classes across the three source languages. Next, the NE tags of target language are projected based on the candidates collected from the three source languages. For the first version, our *strict-match* projection algorithm 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 (Nothman et al., 2013). *Strict match* requires at least a two alignment agreement (the same NE class for an aligned token in two of three source languages) in order to project a NE tag. If there is a tie (usually when multiple alignments occur), then we back-off and project the most frequent NE class in Wikiner for the given token. More specifically:

1. Following the same procedure as for the upper-bound method, we obtain the word alignments for all four languages and order them following the alignment types presented in Table 1.

2. The semantic tags are projected via word alignments from the three source languages to a given fourth, target language. In this step a NE tag is projected following three criteria: (i) if there are at least two agreements between the alignments in the source; (ii) if there is a tie, e.g., if more than two agreements occur, then via back-off; and (iii) if there is not agreement in the alignments, via back-off. This is illustrated by Table 3.

3. This step is the same for the upper-bound method: we assign the span to the projected NE tags in the target assuming that they must be contiguous.

As mentioned in Section 2., we use the corpus Wikiner (Nothman et al., 2013) to back-off whenever *strict match* cannot decide which tag to project. The back-off strategy is fairly simple. Given a large corpus annotated with Named Entities, we calculate the most frequent tag for each token in the corpus. Thus, whenever we need to back-off, we simply consult the frequencies table obtained from Wikiner for the candidate token and assign to that token the most frequent one. For example, the token *European* mentioned in Table 3 could conceivably be LOC, ORG or PER. As in Wikiner the most frequent tag is ORG, then when backing-off that is the tag that will be assigned to that specific token.

4. Experiments

First we present the results of the *upper-bound* projections using the Europarl gold-standard described in section 2. Table 4 displays the overview results of projecting with both 1-1 and multiple alignments. It is clear that the results obtained projecting multiple alignments are better than those with 1-1 alignment for all four languages. This is probably due to the fact that in 1-1 alignments many projections are not performed because no agreement is found.

With respect to the results of the *strict match* projections, the models obtained from the gold-standard data are those described in (Agerri and Rigau, 2016). For training the models on the projected data we induced the same clustering features described in (Agerri and Rigau, 2016) but using the unlabelled Europarl training set instead of the datasets originally used. These clustering features replaced the features used in the original, gold-standard trained models. Thus, the difference between the gold-standard and the projected models mentioned in Tables 5, 6 and 7 is the training corpus (CoNLL-Evalita vs Europarl) and the clustering lexicons used (Wikipedia-Gigaword, etc. vs Europarl).

As we have already mentioned, we compare the gold-trained models with the automatically induced ones via

Alignments	**Tags in projection es; de; it**	**Projected tag**
3-agreement	ORG; ORG; ORG	ORG
2-agreement	ORG; ORG; PER	ORG
no agreement	ORG; LOC; PER	back-off
> 2-agreement	ORG; ORG, LOC; LOC	back-off

Table 3: Strict match projection for the token *European* for English as target language.

Alignment	en	de	it	es
1-1	91.47	75.52	91.75	96.32
Multiple	96.01	94.21	93.50	97.34

Table 4: F1 results on upper-bound projection.

strict match projections. 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 5 show that the automatically trained models outperform the models trained on gold-standard data except for Italian.

Training	en	de	it	es
Gold	65.08	49.87	65.82	58.75
Projected	69.14	70.62	62.44	64.16
upper-bound	96.01	94.21	93.50	97.34

Table 5: Comparing F1 results training ixa-pipe-nerc on projected and gold-standard data.

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 confident. Thus, for Italian the precision results are 7 points higher, 25 points for Spanish and 10 points for German, as it can be seen Tables 6 and 7.

	Precision	Recall	F1
en	70.00	60.34	64.81
de	68.40	39.24	49.87
it	67.03	62.45	64.66
es	55.66	59.69	57.60

Table 6: Evaluating Gold-standard trained CoNLL and Evalita models on Europarl test.

Still, and even though our first results are quite promising, we should note that the results of the automatically generated models are much lower than those established by the upper-bound.

5. Related Work

Traditionally, there are many studies and works exploring the contribution of semantic information or features with the aim of improving Machine Translation (Koehn, 2010; Artetxe et al., 2015) but the reverse has been rather uncommon. Among previous works using parallel texts to automatically induce linguistic processors, most of them focus on inducing Part of Speech taggers (Yarowsky et al., 2001; Ganchev and Das, 2013; Täckström et al., 2012; Fossum and Abney, 2005) although a very few of them worked on semantic tasks such as Named Entity Recognition (NER) (Yarowsky et al., 2001; Zhang et al., 2016) and Semantic Role Labelling (SRL) (Padó and Lapata, 2009).

	Precision	Recall	F1
en	70.30	68.01	69.14
de	78.87	63.94	70.62
it	75.14	53.41	62.44
es	80.29	53.42	64.16

Table 7: Evaluating models trained on automatically projected data.

Furthermore, almost every previous approach is based on one-to-one projections using only one language pair to induce the linguistic processors. As far as we know, there are only two exceptions: Yarowsky et al. (2001) use bridging between two languages to perform lemmatization in a third target language, and Fossum and Abney (2005) train multiple POS taggers from monolingual source data and combine their annotations to project them to a given target language. Therefore, to our knowledge, no previous approach aims at doing transfer of semantic annotations as we propose in this paper.

These previous works based on projection of annotations have shown that the projected labels can result in a very noisy training set in the target language. Various methods have been applied to address this problem, including smoothing techniques (Yarowsky et al., 2001) and the combination of token-level and type-level constraints to recalculate the probability distribution of the labels in a CRF for Part of Speech tagging (Täckström et al., 2012). (Das and Petrov, 2011) use the projected labels as contraints in a Posterior Regularization framework and (Ganchev and Das, 2013) extend this work by training directly discriminative models via cross lingual projection with Posterior Regularization. Finally, instead of using total counts of labels of a class to enforce the constraints, (Wang and Manning, 2014) define expectation constraints at token level for NERC.

Closest to our work, Zhang et al. (2016) generate a high-confidence annotation set using strict rules on parallel corpora in order to project the Named Entity information from the source to the target. The resulting annotated bitext is then used to train a LSTM model. They evaluate their work with respect to a baseline consisting of the projected tags via automatic word alignments. The results show that the word alignment method is much worse than the bitext trained LSTM model. It should be noted that they do not explain how the annotations are projected via word alignments. Furthermore, we believe that using only one source language may be detrimental to the quality of the projections.

6. Concluding Remarks

We train the same tagger on the automatically projected 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. In this paper we have demonstrated that it is feasible to automatically induce training data using parallel data without manual intervention. This method allows to generate Named Entity Recognition (NER) taggers for a given language when no manually data is available. Furthermore, our method may be applied to generate annotations for other semantic tasks, such as Semantic Role Labeling or Supersense tagging.

We believe that the reported results could be improved by using several strategies: First, it could be worth to perform another iteration of the *strict match* projection. After all, the projected NE tags are automatically obtained by applying the gold-standard trained models, which, as shown by the evaluation, are much worse than the models obtained from the Europarl. Thus, tagging the Europarl training set with the automatically obtained models and project those annotations could improve the quality of the projections.

Second, we may include more languages to improve the quality of projections. In our experiments we have considered four languages, three source and one target, but it might be worth to investigate if integrating more source annotations is likely to substantially cancel out projection errors.

Third, we could focus the projection via word alignments by language groups, namely, grouping Romance languages, Slavic languages, under the assumption that word alignments for closely related languages may be of higher quality.

Future work also includes evaluating both gold-trained and projected models on out-of-domain data. 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.

The gold-standard and automatically generated models[1] as well as the the manually-annotated test set from the Europarl[2] are made publicly available for its use and to facilitate reproducibility of results.

7. Acknowledgements

This work has been funded by the Spanish Ministry of Economy and Competitiveness (MINECO/FEDER, UE), under the projects CROSSTEXT (TIN2015-72646-EXP) and TUNER (TIN2015-65308-C5-1-R).

[1] https://github.com/ixa-ehu/ixa-pipe-nerc

[2] https://github.com/ixa-ehu/ner-evaluation-corpus-europarl

8. Bibliographical References

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

Artetxe, M., Agirre, E., Alegria, I., and Labaka, G. (2015). Analyzing english-spanish named-entity enhanced machine translation. In *SSST@ NAACL-HLT*, pages 52–54.

Das, D. and Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 600–609. Association for Computational Linguistics.

Fossum, V. and Abney, S. (2005). Automatically inducing a part-of-speech tagger by projecting from multiple source languages across aligned corpora. *Lecture notes in computer science*, 3651:862.

Ganchev, K. and Das, D. (2013). Cross-lingual discriminative learning of sequence models with posterior regularization. In *EMNLP*, pages 1996–2006.

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

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Liang, P., Taskar, B., and Klein, D. (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.

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

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

Padó, S. and Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36(1):307–340.

Speranza, M. (2009). The named entity recognition task at evalita 2009. In *Proceedings of the Workshop Evalita*.

Täckström, O., McDonald, R., and Uszkoreit, J. (2012). Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 477–487. Association for Computational Linguistics.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147.

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.

Wang, M. and Manning, C. D. (2014). Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association of Computational Linguistics*, 2:55–66.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.

Zhang, D., Zhang, B., Pan, X., Feng, X., Ji, H., and Weiran, X. (2016). Bitext name tagging for cross-lingual entity annotation projection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 461–470.