Building Multilingual Named Entity Annotated Corpora exploiting Parallel Corpora

Maud Ehrmann, Marco Turchi

European Commission - Joint Research Centre (JRC), IPSC - GlobSec,
Via Fermi 2749, 21020 Ispra (VA) - Italy
E-mail: first name.surname@jrc.ec.europa.eu

Abstract

This paper reports first experiments in the automatic building of multilingual named entity annotated corpora, taking advantage of a multiparallel corpus. We believe that providing such a resource could help to overcome the annotated data shortage in the Named Entity field and will guarantee comparability of named entity recognition system results across languages. Our approach is based on annotation projection, which is carried out with the help of a phrase-based statistical machine translation system. We obtain promising results and thus consider proceeding with other languages.

1 Introduction

Named Entity recognition is a well-established task: specified for the first time during the latest American MUC conferences, it is now acknowledged as a fundamental task to a wide variety of natural language processing (NLP) applications. Rule-based, machine learning and hybrid named entity recognition systems have been developed over the years, achieving respectable performances for various languages, domains and applications (Nadeau et al. [14]). As for many other NLP tasks, annotated corpora constitute a crucial and constant need for named entity recognition (NER). Within a development or training framework, annotated corpora are used as models from which machine learning systems (or computational linguists) can infer rules and decision criteria; within an evaluation framework, they are used as a gold standard to assess systems' performances and help to guide their quality improvement, *e.g.* via non-regression tests.

During the last decade, several named entity (NE) annotated corpora were built, thanks to a large series of evaluation campaigns (Fort et al. [7]). However, such resources remain rather rare and limited to a relatively small set of languages and domains. Even if unsupervised methods tried to overcome this difficulty, the shortage of annotated data for the large majority of world's languages remains a problem.

An obvious solution is to manually produce annotated corpora, but it is a complex and time-consuming task and it may be difficult to find experts in specific language.

Beyond annotated corpora's scarcity, another issue lies in the fact that annotation schemas or guidelines usually differ from one annotated corpus to another: named entity extents can be different (*e.g.* inclusion or not of the function in a person name, *Secretary of State Hillary Clinton* vs. *Hillary Clinton*), as well as entity types and granularity (*e.g.* some corpora may consider product names, whereas others will differentiate, within this category, vehicles, awards and documents, and others won't even consider product names). Such divergences should be expected, as annotated corpora are built within different frameworks and according to different applications. However, they constitute a real issue, particularly when developing or evaluating multilingual NE recognition systems. Actually, in a multilingual environment, if someone wants to use named entity annotated corpora (if available), he/she should first convert the data to a common annotation schema and document format before exploiting it. To avoid the annotation schema conversion step, Bering et al. [3] built a flexible evaluation tool; although efficient, this solution seems quite heavy to implement and requires a meticulous study of the different annotation schemas.

Our goal is to automatically build a set of multilingual named entity annotated corpora, taking advantage of the existence of parallel corpora (multiparallel or bilingual). Traditionally used in the field of Machine Translation, parallel corpora have been exploited in recent years in various NLP tasks, including linguistic annotation, with the creation of annotated corpora. The underlining principle is *annotation projection*, where annotations available for a text in one language can be projected, thanks to the alignment, to the corresponding text in another language, creating herewith a newly annotated corpus for a new language.

This method shows several advantages. Firstly it could be a way of overcoming NE annotated data shortage problem. Then, it could solve the non-harmonized annotation issue: if the projected annotations (on the target side) always come from the same automatic recognition system (on the source side), then we obtain annotated corpora in different languages, but with a common annotation schema. The use of multiparallel corpora also presents the benefit of ensuring the comparability of NER system results across languages; morever, as named entity recognition systems are domain-sensitive, it could be relevant to evaluate multilingual NER systems on equivalent tasks.

This paper relates our first attempt to apply this method to Named Entity annotations, projecting automatically annotated English entities to French, Spanish, German and Czech aligned corpora. Following this preliminary work, our objective is to automatically annotate and make freely available named entity corpora in a large set of languages, with a quality similar to that of manually annotated data.

The remainder of the paper is organized as follows. In section 2 we introduce related work; we then present our NE projection method (still at its first stage of development) in section 3, report the results in section 4 and finally conclude and propose some elements for future work in section 5.

2 Related Work

Regarding the automatic acquisition/building of NE annotated corpora, some work investigate how to constitute monolingual annotated data: An et al. [1] extract a huge amount of documents in Korean from the web and then annotate them automatically whereas Nothman et al. [15] make use of Wikipedia to create a named entity annotated corpus in English, transforming Wikipedia's links into NE annotations. In each case, the resulting corpora allow the authors to train a NER system that performs quite well, thus vouching for the newly labeled data quality.

With regard to parallel corpora, their exploitation has been growing in recent years, showing their usefulness in various NLP tasks like word sense disambiguation or cross-lingual tagging (refer to the state of art presented by Bentivogli et al. [2]). With respect to cross-lingual knowledge induction, multiple work addressed the challenge of automatic parallel treebank building, deducing syntactic information correspondences (Lavie et al. [12]) or projecting them from one language to another (Hwa et al. [8]). In addition, recent work carried out semantic information projection, mainly focusing on semantic roles and word senses (Padó et al. [16] and Bentivogli et al. [2]).

Several researchers investigated named entity annotation and parallel corpora exploitation. Klementiev et al. [9] proposed an algorithm for cross-lingual multiword NE discovery in a bilingual weakly temporally aligned corpus. Their goal is to extract pairs of named entities across languages, by co-ranking two clues: synchronicity (use of a time distribution metric) and phonetical similarity (use of a transliteration model). Ma [13] applies a co-training algorithm on unlabelled bilingual data (English-Chinese), showing that NE taggers can complement and improve each other while working together on parallel corpora. Samy et al. [17] developed a named entity recognizer for Arabic, leveraging an Arabic-Spanish parallel corpus aligned at sentence level and POS tagged. Yarowsky et al. [21] achieved some pioneer experiments, exploring the feasibility of annotation projection in four tasks, one of which was named entity annotation. The goal was to automatically induce stand-alone text analysis tools via robust annotation projection. Such approaches deal with named entity annotation and make use of parallel corpora but mainly aim at developing or improving NER systems; it seems that parallel annotated corpora are a positive side-effect of these work, but they don't go into details. Our approach differs in that we focus our attention on acquiring multilingual annotated corpora mainly for evaluation purpose. Therefore, high precision is required and we cannot afford noisy projections.

Finally, the work of Volk et al. [20] on combining parallel treebanks and geotagging offers similar results to what we propose, with the difference that they focus on the *location* type, ground the annotated entities with references to a gazetteer and work with a bilingual French-German corpus.

3 Named Entity Annotation projection

Given a multiparallel corpus and a monolingual NER system, our objective is to automatically provide NE annotations for each text of the aligned corpus. We assume that a possible solution to project a named entity from a text in one language to an aligned text in another language is to translate this entity, using different approaches, e.g. machine translation. Following this assumption, our multilingual NE annotation projection method relies, for the most part, on the use of a phrase-based statistical machine translation system (PBSMT). We used a multiparallel corpus in English, French, Spanish, German and Czech, that is news texts coming from the WMT shared tasks (Callison-Burch et al. [5]). For each language, we have a training set of roughly 70,000 sentence pairs and a test set of 2,490 sentence pairs. We used the test set for the annotation projection. The next sections detail each step of the NE annotation projection process.

3.1 Automatic annotation of source Named Entities

The first step is to annotate NEs in one corpus in a given language. We chose to annotate English entities of type *Person*, *Location* and *Organisation* and tried to project them in the corresponding texts in other languages. As a matter of fact English is a resource-rich language with already existing efficient tools, but one may choose another source language, according to his/her own goals and constraints.

We used an in-house NER system (Steinberger et al. [18] and Crawley [6]) to process the English source side text (any NER system or even manual annotation could have been used at this stage). It is obvious that the NER system quality is a crucial element that determines the projection quality: if the system misses one entity or wrongly annotates it, it won't be projected or it will be wrongly annotated. In our English text, the NER system annotated a total of 826 unique entities, corresponding to 1,395 entity occurrences, among them 649 person names, 412 location names and 332 organisation names.[1]

3.2 Source Named Entity translation

The second step corresponds to the translation of the previously extracted entities into French, Spanish, German and Czech. We firstly present the Phrase-Based Statistical Machine Translation system and account for its benefits in this particular task; we then report a correction phase and an evaluation of the NE translation.

Phrase-Based Statistical Machine Translation System. One of the most popular classes of statistical machine translation (SMT) systems is the Phrase Based Model [11]. It is an extension of the noisy channel model, introduced by [4], using phrases rather than words. A source sentence f is segmented into a sequence of I phrases $f^I = \{f_1, f_2, \ldots f_I\}$ and the same is done for the target sentence e,

[1] In this paper we do not go into details regarding the source NE annotation (type granularity, extents, etc.) as we focus more on the validation of the approach.

where the notion of phrase is not related to any grammatical assumption; a phrase is an n-gram. The best translation $\hat{e}$ of f is obtained by:

$$\hat{e} = arg\max_{e} p(e|f) = arg\max_{e} \prod_{i=1}^{I} \phi(f_i|e_i)^{\lambda_\phi} d(a_i - b_{i-1})^{\lambda_d} \prod_{i=1}^{|e|} lm(e_i|e_1 \ldots e_{i-1})^{\lambda_{lm}}$$

where $\phi(f_i|e_i)$ is the probability of translating a phrase e_i into a phrase f_i. $d(a_i - b_{i-1})$ is the distance-based reordering model that drives the system to penalize significant reordering of words during translation, while still allowing some flexibility. In the reordering model, a_i denotes the start position of the source phrase that was translated into the ith target phrase, and b_{i-1} denotes the end position of the source phrase translated into the $(i-1)$th target phrase. $lm(e_i|e_1 \ldots e_{i-1})$ is the language model probability that is based on the Markov chain assumption. It assigns a higher probability to fluent/grammatical sentences. λ_ϕ, λ_{lm} and λ_d are used to give a different weight to each element. For more details see [11].

Phrases and probabilities are estimated processing the parallel data. Word to word alignment is firstly extracted running the IBM models [4], and then, on top of it, proximity rules are applied to obtain phrases, see [11]. Probabilities are estimated counting the frequency of the phrases in the parallel corpus. In this work, we used the PBSMT system Moses [10].

Among all the possible translation techniques, we decided to use this approach because, in general, entities are a small set of contiguous words, phrases, and PBSMT systems perform better than systems based on single words. In this work, we do not apply the classical idea of translation: a sentence that is not present in the training data (unseen sentence) is translated to another language. In our experimental framework, we train a PBSMT system using as training data the parallel sentences that we want to annotate plus a larger set of parallel sentences. This means that the translation system knows how to translate the source entity, because it has seen it in the training data; this reduces the number of completely untranslated entities. At the end, we use the SMT system for its capability of aligning bilingual phrases across two parallel sentences more than for its capability of translating unseen sentences. Unfortunately, this experimental setting does not guarantee that all the source entities are always correctly translated, because its statistical approach favours those translations that appear more often in the training data. That's why we added a correction phase after the translation.

Correction phase. Entity translations are not always correct because the PBSMT system tries to reproduce the most readable sentence driven by the language model; in this way, the translation system may add articles, prepositions or in some cases groups of words before or after the entity name. For example, the french translation of *Afghanistan* is *en Afghanistan* and the translation of *Germany* is *l'Allemagne*. In these cases, only *Afghanistan* and *Germany* should be projected, as prepositions and articles cannot be part of proper names in French. We could observe similar phenomena in other languages.

To address this problem, we post-processed the translations in a simple way: applying stopword lists. This allowed us to correct a certain number of entities

for each language, even if some wrong entities could remain in the list. Before projecting these "corrected" translated entities in the aligned corpora, we asked bilingual annotators to check the correctness of the translated entities.

Evaluation of the NE translation. We randomly selected two hundred English entities and their relative translations in French, Spanish, German and Czech. We then provided annotators with the bilingual lists plus a set of evaluation categories that identify possible translation errors:

1. Correct Translation: the translated entity is correctly translated.
2. Extra Words: the translated entity contains some superfluous words (En: *tariq ramadan* Fr: *peut-être tariq ramadan*).
3. Missing Words: the translated entity does not contain some original words (En: *eastern punjab* Fr: *punjab*).
4. Wrongly Translated Words: the translated entity contains some words that are incorrectly translated (En: *reuters news agency* Fr: *nouvelle agence reuters*).
5. Wrong Word Order: some words in the translated entity are not correctly located (En: *south africa* Fr: *sud du afrique*).
6. Wrong Translation: the translated entity is wrong.

Evaluation results are reported in Table 1. In all languages, main problems seem to be the addition and subtraction of word(s) during the translation phase. This comes from the fact that the PBSMT tries to reproduce the most readable sentence (as pointed above), adding or removing some words that afterwards were not removed by the stoplists. We also observe that there are more completely wrong translations when French or Spanish are the target language. Presumably, this is due to the fact that there are different translation choices (verbatim or not) between languages for specific names such as *Canada Cup*, *Stanley Cup* or *Walmart Foundation*; in front of this situation, the annotators adopted different behaviours. We need to investigate this phenomenon, in order to know if we can predict when it is preferable not to translate, but to keep the English entity.

In general, SMT performance depends on the training set size [19]. We first trained the PBSMT system with the parallel sentences that we want to use in the projection only, obtaining poor results. For this reason, we then added more training data, whose size (70,000 sentence pairs) is still rather small according to the machine translation community standards. We believe that adding more data can increase the translation performance and in particular solve the problem of unwanted or deleted words in the translations.

	French	Spanish	German	Czech
Correctly translated	83,5	83.5	82.5	83.5
Extra words	4.0	3.0	7.0	9.0
Missing words	3.0	4.5	6.5	3.5
Wrong words	2.0	1.0	0.0	1.0
Wrong order	1.0	0.5	0.5	0.5
Wrong translation	6.5	7.5	3.5	2.5

Table 1: Human Evaluation of NE translation (error type percentages).

3.3 External Named Entity resource

In addition to the SMT approach, we benefit from an external multilingual named entity resource. The JRC's named entity database has been built up since 2004 through a daily analysis of tens of thousands of multilingual news articles per day; it contains, among others, translations and transliterations of entity names in several languages [18]. By querying this database, we retrieved, for each English entity, a list of translated entities (that may have different spellings) in a given language.[2]

The information coming from the external resource is quite reliable, because part of the entity names has been manually checked. However, it is not exhaustive. On the contrary, the SMT system provides translations almost every time, but they may be incorrect. In other words, information coming from the external resource and the SMT system can complement each other, the former boosting precision and the latter ensuring recall. For example, *Sakharov Prize for Freedom of Thought* is correctly translated by the SMT system for each language while the database does not contain this name.

3.4 Named Entity projection

Once we have a list of possible translations for a given NE in a particular sentence, we try to project it into the corresponding sentences of the aligned corpora, using a simple and strict string matching: the translation is present or not. We applied the whole processing chain to our multiparallel corpus; the next section presents the projection results.

4 Results and discussion

We evaluate the performance of the projection using three different translation approaches. English entities are translated using: (1) external information: for each

[2]The database contains 134,046 en-fr named entity translations, 157,442 en-sp, 156,363 en-de and 2,807 en-cs.

language pair, a list of English-Foreign entity associations is used as a look-up table (*Ext* in Table 2), (2) machine translation system (*SMT*) and (3) external information and machine translation system together: a list of all possible translations is associated to each English entity[3](*All*).

As we do not have a reference corpus, we can only compute projection's Recall. An indirect way to evaluate the Precision is the SMT evaluation, but this is only a partial evaluation. In the future, we will vary our projection method (not only strict string matching) and manually annotate a part of the multilingual set to provide a complete evaluation of the projection.

During the first step (source NE annotation), we noticed the presence of wrong English entities. In this work, we do not evaluate the quality of the NER system that we used, but we are interested in evaluating how it affects our projection performance. For this purpose, we manually corrected the English entities. In Table 2, we report results for projections done using all the English entities and only the correct ones. Recall is computed relative to the total number of English entities.

	French	Spanish	German	Czech
Ext	0.325	0.264	0.291	0.103
Ext (En Correct)	0.343	0.278	0.306	0.106
SMT	0.798	0.787	0.794	0.535
SMT (En Correct)	0.825	0.806	0.813	0.545
All	0.807	0.800	0.807	0.547
All (En Correct)	0.834	0.821	0.827	0.557

Table 2: Recall of the annotation projection.

The first observation is that projections are strongly affected by the target language. When French, Spanish and German are the target languages, performances are similar, while with Czech there is a drastic drop in performance. This is due to the fact that Czech is a highly inflected language and for the same English entity there are more than one possible translation (morphological variants).

Projections obtained using only the external resource produce low recall. This approach is quite good for those English entities that have a standard form like first name-surname (e.g. *Matt Damon*) or location names (e.g. *South Africa*), but is less efficient for organization entities (e.g. *Czech hydrometeorological institute*). The big advantage of using an SMT system trained with the data that we want to use during the projection is that all the information is available for the SMT system which can correctly translate entities, even complex ones. This aspect can be seen in the results, where recall with SMT translation improves substantially compared to the recall obtained using the external resource only. Merging of external and SMT translations produces small improvements, while removal of wrong English entities affects positively the results, in particular for German, Spanish and French.

[3]If more than one translation matches the target sentence, it is counted only one time.

5 Conclusion and Future Work

Parallel corpora can support the automatic creation of multilingual NE annotated corpora. We presented preliminary experiments of a NE annotation projection method for a 5 language multiparallel corpus, obtaining encouraging results.

The current approach can be improved in several ways. First of all, as demonstrated by different results with/without wrong English entities, we need to improve the NER system. Then the projection approach (presence/absence of the translated entity) is particularly strict. We believe that different methods based on word similarity and word alignment can be used to find the correct entity in the target sentence.The main issue is the projection of the entities in a highly inflected language. To solve this problem, one solution is to force the PBSMT system to emit also the less probable translations, trying to cover all possible variations in the inflected language. Finally, we intend to apply this method to other parallel corpora in different languages.

Acknowledgements We would like to thank Josef Steinberger and Ralf Steinberger for accepting to annotate Czech and german entities, as well as J. Brett Crawley, Jenya Belyaeva, Vanni Zavarella and again Ralf for their comments.

References

[1] An, J., Lee, S. and Lee, G. (2003) Automatic acquisition of named entity tagged corpus from world wide web. In *Proceedings of the 41st Annual Meeting on ACL (ACL'03)*, Sapporo.

[2] Bentivogli, L. and Pianta, E. (2005) Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. In *Natural Language Engineering* pp. 247–261, Cambridge University Press.

[3] Bering, C., Drozdzynski, W., Erbach, G., Guasch, C., Homola and others. (2003) Corpora and evaluation tools for multilingual NE grammar development. In *Proceedings of Multilingual Corpora - Linguistic Requirements and Technical Perspectives*, Lancaster.

[4] Brown, P.F., Della Pietra, S., Della Pietra, V.J. and Mercer R.L.(1994). The Mathematic of Statistical Machine Translation: Parameter Estimation. In *Computational Linguistics*, 19(2):263–311.

[5] Callison-Burch, C., Koehn, P., Monz, C. and Schroeder, J. (2009) Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth WMT'09*, Athens.

[6] Crawley, J. B. and Wagner, G. (2010). Desktop text mining for law enforcement. In *Proceedings of ISI'10*, Vancouver.

[7] Fort, K., Ehrmann M. and Nazarenko, A. (2009) Towards a Methodology for Named Entities Annotation. In *Proceedings of LAWIII*, Singapore.

[8] Hwa, R., Resnik, P., *et al*. (2005). Bootstrapping parsers via syntactic projection across parallel texts. In *Natural Language Engineering*, 11(3).

[9] Klementiev, A. and Roth, D. Named Entity Transliteration and Discovery from Multilingual Corpora. (2008) In *Learning Machine Translation.* MIT Press.

[10] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N. and others (2007). Moses: Open source toolkit for statistical machine translation. ACL, 45(2), Columbus, Oh, USA.

[11] Koehn, P. (2010). Statistical Machine Translation. Cambridge Univ. Press.

[12] Lavie, A., Parlikar, A. and Ambati, V. (2008). Syntax-driven learning of subsentential translation equivalents and translation rules from parsed parallel corpora. In *Proceedings of theHLT-SSST-2 workshop*, Columbus, Ohio.

[13] Ma, X. (2010) Toward a Name Entity Aligned Bilingual Corpus. In *Proceedings of the Seventh LREC Conference*, Malta.

[14] Nadeau, D., and Sekine, S. (2007) A survey of named entity recognition and classification. In *Linguisticae Investigaciones*, 30-1, pp. 3-26.

[15] Nothman, J., Curran, J., and Murphy, T. (2008) Transforming Wikipedia into named entity training data. In *Proceedings of the ALTA Workshop*, Hobart.

[16] Padó, S. and Lapata, M. (2009) Cross-linguistic projection of role-semantic information. In *Journal of Artificial Intelligence Research*, 36.

[17] Samy, D., Moreno-Sandoval, A. and Guirao, J.M. (2005). A Proposal for an Arabic Named Entity Tagger Leveraging a Parallel Corpus (Spanish-Arabic). In *Proceedings of RANLP Conference*, Borovets, Bulgaria.

[18] Cross-lingual Named Entity Recognition. Steinberger, R. and Pouliquen B. (2007). In *Linguisticae Investigationes*, 30-1, John Benjamins.

[19] Turchi, M., DeBie, T. and Cristianini N. (2008). Learning Performance of a Machine Translation System: a Statistical and Computational Analysis. In *Proceedings of the Third WMT'08*, Columbus, Oh, USA.

[20] Volk, M., Goehring, A. and Marek, T. (2010) Combining Parallel Treebanks and Geo-Tagging. In *Proceedings of The Fourth LAW Workshop*, Uppsala.

[21] Yarowsky, D., Ngai, G. and Wicentowski, R. (2001) Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of HLT'01*, San Diego.