Cross-lingual Named Entity Recognition via FastAlign: a Case Study

Ali Hatami[1], Ruslan Mitkov[1], and Gloria Corpas[2]

[1] University of Wolverhampton, UK
{a.hatami, r.mitkov}@wlv.ac.uk
[2] University of Málaga, Spain
gcorpas@uma.es

Abstract. Named Entity Recognition is an essential task in natural language processing to detect entities and classify them into predetermined categories. An entity is a meaningful word, or phrase that refers to proper nouns. Named Entities play an important role in different NLP tasks such as Information Extraction, Question Answering and Machine Translation. In Machine Translation, named entities often cause translation failures regardless of local context, affecting the output quality of translation. Annotating named entities is a time-consuming and expensive process especially for low-resource languages. One solution for this problem is to use word alignment methods in bilingual parallel corpora in which just one side has been annotated. The goal is to extract named entities in the target language by using the annotated corpus of the source language. In this paper, we compare the performance of two alignment methods, Grow-diag-final-and and Intersect Symmetrisation heuristics, to exploit the annotation projection of English-Brazilian Portuguese bilingual corpus to detect named entities in Brazilian Portuguese. A NER model that is trained on annotated data extracted from the alignment methods, is used to evaluate the performance of aligners. Experimental results show the Intersect Symmetrisation is able to achieve superior performance scores compared to the Grow-diag-final-and heuristic in Brazilian Portuguese.

Keywords: Named Entity Recognition · Word Alignment · Cross-lingual.

1 Introduction

Word alignment is a Natural Language Processing (NLP) task that can be applied to a parallel text corpus to find the word-to-word correspondences in a sentence pair. It can be used in the Machine Translation (MT) pipeline to analyse the output of MT or to improve the quality of translation memory [1]. Although word alignment is not a necessary step in the MT pipeline, in special cases it helps the model to improve the translations. For example, word alignment is useful to translate domain-specific terminology and low-frequency content words [2]. Word alignment can also be used to match the alignments with the source annotations to determine the projection on the target text. It helps the

Translation and Interpreting Technology Online, pages 85–92,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_010

Named Entity Recognition (NER) system to detect Named Entities (NEs) on the target text based on the annotation information of the source text.

NER is a classification task that tries to identify and tag NEs in a given sentence. The benefits of performing NER during translation tasks range from improved translation quality to customer data protection. The performance of a machine learning-based NER system depends on multiple factors such as the amount of labelled data used to train the model. The availability of labelled data is one of the key points for training the NER systems. It is a basic challenge, especially for low-resource languages. One of the possible solutions for this challenge is to use cross-lingual approaches. These approaches use parallel bilingual corpus to extract the annotation information from languages with rich labelled data.

There are two different approaches to extract NEs from bilingual corpora: direct transfer approach, and annotation projection approach. Approaches based on the direct transfer try to use language-independent features to train the model on the source side and then directly apply it on the target side. Different cross-lingual features such as word-embeddings [3], word clusters [4] and Wikifier [5] can be used in the training process. These approaches suffer from sense ambiguity and word order differences that lead to noise in the output of the model [6].

Methods based on the annotation projection use word alignment information to project annotations from the source language to the target side in bilingual parallel corpora. Different approaches can be used to extract the alignments between sequences of words in the source and target languages. The most common approaches of word alignments are based on the IBM approach that is a classic word alignment model [7]. In the IBM models, every word in the source language can be aligned at most with one word in the target language that is called many-to-one mapping. But real-word alignment approaches that are based on the IBM model support different types of mappings such as many-to-many. The simplest approach to produce a many-to-many mapping is the symmetrization heuristic [8]. The symmetrization heuristic uses alignment in both directions. There are different methods to implement a symmetrization heuristic. Intersection and Union alignments are the most popular methods that merge alignments of both directions. The Intersection of two models expresses just a one-to-one relationship between words and it misses some of the alignments. So, it has a higher precision of alignment points but at the cost of losing in recall. Union of two models can capture all complementary information of both models. Unlike the Intersection heuristic, Union has a higher recall but lower precision.

There are some extended versions for Intersection to improve its performance such as the GDFA heuristic. It grows the Intersection heuristic by adding neighbouring alignment points from the union and unaligned points to the intersection [9]. The GDFA heuristic includes three steps. The first step is Grow-diag that intersects two-directional alignments and gradually considers the neighbourhood of each alignment point between the source and target languages. The second step is (-final) covers non-neighbour alignment points of intersection alignment points. The final step (-and) adds alignment points between two unaligned words.

This paper is organized as follows. Section 2 introduces NER and alignment methods. Section 3 describes experimental settings and results. Finally, Section 4 presents our conclusions and future work.

2 Methods

Word alignment provides useful information for several applications of NLP. MT is one of these applications that can use the benefits of word alignment methods to improve translations. Although the Neural MT models (NMT) do not rely on the word alignment approaches, they still play an important role to improve the output quality of the model. They can be used to extract an external lexicon and apply it in the inference process of MT [10]. It can help the MT model to better use domain-specific terminology to adapt the model with a new domain or to improve the translations of out-of-vocabulary content words. NER is another application to use word alignments in MT to improve its performance. NER is an information extraction task to automatically detect NEs in text and classifying them into predefined entity types such as PERSON, ORGANIZATION, LOCATION, TIME, DATE, etc. Today's NER systems are based on supervised machine learning models including Maximum Entropy Markov Models (MEMMs) [11], Conditional Random Fields (CRFs) [12], and neural networks [13]. Although the NER model that is based on neural architectures show high performance, they need a large amount of manually annotated NER data that is not available for low-resource languages [14]. In addition, the process of annotating by a human is a time- and money-consuming task.

Cross-lingual NER is an effective solution to tackle these challenges. It means transferring annotated information from a high-resource language that has enough annotated resources to a low-resource language with less or no annotated data. There are two main groups of cross-lingual NLP, direct transfer and annotation projection [6]. In direct transfer approaches, the NER model is trained on a language with rich labelled data and then applied to a text in a different language to detect NEs [4]. These approaches attempt to use language-independent features. But some features are dependent on the type of language. So, selecting a suitable set of features plays an important role in the quality of these approaches.

Annotation projection is another approach that is based on a parallel corpus between source and target languages [15]. These methods attempt to annotate the target side by using the annotation information of the source language. The quality of the annotation task is related to the quality of labelled data on the source side, the quality of alignments, and the size of the parallel data. This paper is a part of our research on cross-lingual NER transfer with minimal resources in the pipeline of MT. We focus on evaluating the performance of alignment methods in the annotation projection approach, where there is only one source language with rich label and no labelled data in the target language. In this paper, the accuracy of two alignment heuristics, Grow-diag-final-and (GDFA) and Intersect Symmetrisation are evaluated on the English-Brazilian Portuguese parallel corpus.

3 Experiments

In this paper, the NER model is based on the FLAIR [1] framework to train it on the training dataset and extract NER annotations from the test dataset. Before training the NER model, we used aligner methods to extract the annotation projection for the target language, Brazilian Portuguese, because we want to figure out ways of taking annotation benefits of the source language, English, that already is available.

3.1 Experimental Settings

The NER model of our experiment is on the FLAIR framework. The main idea of this framework is based on the word and document embeddings. It uses a simple GloVe embedding for 150 epochs to train the model [16]. We trained the model on our datasets (Logitech, Rakuten, TomTom and Udemy) in the English-Brazilian Portuguese language pair. Then performance of the trained model is evaluated on the test dataset. About 10% of the total data has been selected for each set, validation and test. Table. 1 shows the statistical information of training, validation and test datasets for the English-Brazilian Portuguese corpus. The entire dataset including training and test, has been manually annotated in the source side (it is called Gold-standard reference) as follows:

'text': "Use the Windows calibration utility."
$[' token\_end' : 3,\ ' start' : 10,\ ' label' :' PRS',\ ' end' : 17,\ ' token\_start' : 3]$

In this research, the fast-align [2] model has been used to extract correspondence words (or multi-words) for the English-Brazilian Portuguese language pair. The main focus of this part of the project is to extract annotations of the target text by matching the alignments with the source annotations. Parameters of the alignment model were trained based on the generic data. Because the alignment model needs the tokenized source and target sentences as input. For the source side of the test dataset, we have tokenized sentences and manual annotation was provided by a linguistics expert. After tokenizing the target sentences, alignment algorithms were applied to the test dataset for extracting the annotation projection between the source and target languages. For example, in the sentence pair of "Go to Kobo.com. || Vá para Kobo.com.", the term of "Kobo.com" is a named entity with a tagged as URL label. An aligner tries to find correspondence words of the source text in the target text. In this example, the output of aligner is: Go → Vá || to → para || Kobo.com → Kobo.com. So we can detect the named entity of "Kobo.com" using projection between the source and target languages.

The GDFA and Intersect symmetrisation heuristics are used to obtain alignments. These heuristics use different alignment approaches in both directions (EN→BR-PR and BR-PR→EN). One of the notable differences between the outputs of GDFA and Intersect heuristics comes back to the approach of aligning

[1] $https://github.com/flairNLP/flair$
[2] $https://github.com/clab/fast_align$

multi-words to a single word. GDFA heuristic finds all correspondences between words, but Intersect just aligns one of the multiple words (first word) to a single word. For example, in the sentence pair of "Thank you for waiting. || Obrigado por esperar.", the term of "Thank you" in the source language must be aligned to "Obrigado" in the target language. The GDFA heuristic aligns both words, "Thank" and "you" to the target word (0-0 1-0 2-1 3-2 4-3) but the intersect heuristic just aligns "Thank" and omits the second word (0-0 2-1 3-2 4-3).

	Source Language			**Target Language**		
	Training	**Dev.**	**Test**	**Training**	**Dev.**	**Test**
Number of sentences	2811	281	363	2811	281	363
Number of words	21358	3858	5035	42146	4404	5106
Unique words	2251	1174	776	3645	1082	929
Number of NE	1153	112	88	1153	112	88

Table 1. Data statistics for the EN-BR PR corpus.

3.2 Results & Discussion

To evaluate the performance of the aligner algorithms, the accuracy of annotation projection can be investigated. Gold-standard annotation is used as a reference to evaluate the results of the projection. Table. 2 shows the accuracy of each label and overall accuracy for both heuristics based on Gold-standard reference. The accuracy of Intersect heuristic is 75% which is 10% higher than the accuracy of the GDFA heuristic. PRS and ORG are two labels that make this difference between the accuracy of two heuristics. The accuracy of the Intersect heuristic in PRS and ORG tags are 69% and 86%, respectively, while those of GDFA heuristic are 58% and 73%, respectively.

To evaluate the performance of the alignment heuristics, we use the NER model that trained on the annotated dataset. The annotated information for the training dataset can be extracted from the annotation projection by the aligners, GDFA & Intersect. The GDFA & Intersect heuristics were used to extract annotation in the target side by using annotation projection between the source language and the target language. Table. 3 shows the results for the NER model based on GDFA & Intersect heuristics respectively.

For the test dataset, Gold standard reference was used to evaluate the performance of the NER model as well as annotation data provided by heuristics. So, we trained the NER model on corpora that annotated using two different heuristics. For each model, we used a test dataset which was labelled in two different ways, relevant heuristic method and the Golden reference. The results show that the intersect-based NER model has higher f1-score than the NER model which trained on the GDFA heuristic. Based on the results of the alignment experiments (Table. 2), NEs in our test dataset can be grouped into 7 classes

	GDFA			Intersect		
label	**det.**	**no-det.**	**Acc (%)**	**det.**	**no-det.**	**Acc (%)**
NAME	8	2	80	8	2	80
PRS	32	23	58	38	17	69
URL	6	4	60	6	4	60
ORG	11	4	73	13	2	86
REFNUM	1	3	25	1	3	25
EMAIL	0	1	0	0	1	0
CRR	0	1	0	0	1	0
Overal	**58**	**38**	**65**	**66**	**30**	**75**

Table 2. Accuracy of GDFA & Intersect based on the source side's annotation.

including: NAME, PRS, URL, ORG, REFNUMBER, EMAIL and CRR. The NER model for both aligners detected some wrong classes that do not have correct named entity labels. This problem comes back to the setup of the aligners that aligned a named entity into the wrong label on the target side. The output of the aligners without correcting the wrong labels has been used to train the NER model. The results show that the Intersect-based NER model provides a better f1-score than the GDFA-based model in 6 classes out of 7. Only the f1-score for the class of "PRS" in the GDFA-based model (0.4138) is a bit better than that of the Intersect-based (0.4).

	GDFA-based NER		Intersect-based NER	
Reference	**Grow**	**Gold**	**Intersect**	**Gold**
NAME	0.3636	0.3636	0.3636	0.6154
PRS	0.3571	0.4138	0.3860	0.4
URL	0.4	0.8	0.8571	0.8
ORG	0.7619	0.7692	0.7692	0.8889
REFNUM	0.2857	0.3333	0.6667	0.3333
EMAIL	0.667	0.0	0.0	0.0
CRR	0.0	0.0	0.0	0.0
weighted avg.	**0.9788**	**0.9818**	**0.9861**	**0.9865**

Table 3. F1-score of NER model based on GDFA & Intersect.

4 Conclusion & Future Work

During this study, we focused on NER as a crucial task in the machine translation pipeline. The availability of labelled data for training the model is a main challenge of the NER systems. The focus of this project was to address this problem by using the aligner algorithms. The aligners can extract annotations in a target side (low-resource language) using annotation projection from the

source side (high-resources language). The experiment shows the NER model that was trained on the Intersect heuristic has a better performance than GDFA. It seems that the performance of the aligners directly impacts on the performance of the NER model. Using state-of-art approaches for the alignment part can be a potential plan for future projects. In this project, restrictions on access to parallel bilingual dataset in English-Brazilian Portuguese by the source side's annotations impacts on the performance of the NER model as well as aligners. This project can be considered as a starting point for our study on the aligner approaches for annotation projection in low-resource languages.

References

1. Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio: Neural Machine Translation by Jointly Learning to Align and Translate. International Conference on Learning Representations, (2016).
2. Philip Arthur, Graham Neubig, Satoshi Nakamura: Incorporating Discrete Translation Lexicons into Neural Machine Translation. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Pages 1557–1567 (2016).
3. Chen-Tse Tsai, Dan Roth: Cross-lingual wikification using multilingual embeddings. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics. Pages 589–598 (2016).
4. Oscar Täckström, Ryan T. McDonald, and Jakob Uszkoreit.: Cross-lingual word clusters for direct transfer of linguistic structure.Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Pages 477–487 (2012).
5. Chen-Tse Tsai, Stephen Mayhew, Dan Roth: Cross-lingual named entity recognition via wikification. Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. Association for Computational Linguistics, Pages 219–228 (2016).
6. Stephen Mayhew, Chen-Tse Tsai, Dan Roth: Cheap Translation for Cross-Lingual Named Entity Recognition. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Pages 2536–2545 (2017).
7. Chris Dyer, Victor Chahuneau, Noah A. Smith: A Simple, Fast, and Effective Reparameterization of IBM Model 2. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Pages 644–648 (2013).
8. Franz Josef Och, Hermann Ney: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, Volume 29, Number 1, Pages 19–51 (2003).
9. Philipp Koehn, Franz J. Och, Daniel Marcu: Statistical Phrase-Based Translation. Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Pages 127–133 (2003).

10. Stephen Mausam, Soderland, Oren Etzioni, Daniel S Weld, Kobi Reiter, Michael Skinner, Marcus Sammer, Jeff Bilmes, et al.: Panlingual lexical translation via probabilistic inference. Artificial Intelligence, 174(9-10):619–637 (2010).
11. Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira.: Maximum entropy markov models for information extraction and segmentation. In Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pages 591–598 (2000).
12. John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pages 282–289 (2001).
13. Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer: Neural architectures for named entity recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pages 260–270 (2016).
14. Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, Jaime Carbonell: Neural Cross-Lingual Named Entity Recognition with Minimal Resources. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Pages 369–379 (2018).
15. Mengqiu Wang and Christopher D Manning.: Cross-lingual projected expectation regularization for weakly supervised learning. Transactions of the Association for Computational Linguistics, Volume 2, Pages 55–66 (2014).
16. Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, Roland Vollgraf: FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). Association for Computational Linguistics, Pages 54–59 (2019).