eman ta zabal zazu

Universidad del Pais Vasco

Euskal Herriko Unibertsitatea

Automatic Generation of Named Entity Taggers Leveraging Parallel Corpora

Author: Yi-Ling Chung

Advisors: Rodrigo Agerri and German Rigau

Master's Thesis

Master in Language Analysis and Processing
University of the Basque Country (Euskal Herriko Unibertsitatea) UPV/EHU
Department of Computer Systems and Languages

Donostia-San Sebastián, September 2017

Acknowledgments

First of all, I sincerely thank Rodrigo Agerri and German Rigau for offering the opportunity to make this research project feasible. Their enthusiasm towards research and attention to detail have both inspired me to discover interesting insights. Most importantly, I also greatly appreciate their endless patience and detailed explanations to be able to finish the project.

I would also like to thank all the professors from the IXA Research Group, especially Nora Aranberri, Gorka Labaka and Itziar Aldabe for providing their linguistic and technical assistance.

Finally, I am very grateful to all the friends I met in Donostia who have provided any forms of supports to my life and study since the first day I came to this lovely city.

Eskerrik asko!

Abstract

The lack of hand curated data is a major impediment to developing statistical semantic processors for many of the world languages. A major issue of semantic processors in Natural Language Processing (NLP) is that they require manually annotated data to perform accurately. Our work aims to address this issue by leveraging existing annotations and semantic processors from multiple source languages by projecting their annotations via statistical word alignments traditionally used in Machine Translation. Taking the Named Entity Recognition (NER) task as a use case of semantic processing, this work presents 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 et al., 2006). The projected annotations can 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. 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 out-of-domain data. This shows that our projection algorithm is effective to transport NER annotations across languages via parallel data thus providing a fully automatic method to obtain NER taggers for as many as the number of languages aligned via parallel corpora.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Natural language processing (NLP) aims at comprehending and producing human language automatically. NLP covers various linguistic levels including, syntax, semantics, discourse, speech, etc. This work focuses on a very useful semantic task, namely, entity recognition (NER) (Tjong Kim Sang, 2002; Nadeau and Sekine, 2007). The aim of NER is to detect accurately named entities in the text and classify them into general semantic categories, e.g., person, location, organization or date. NER plays an important role in one of the first steps in NLP to capture named expressions appearing in texts. Current state-of-the-art NER systems perform quite well in classifying general categories (Ratinov and Roth, 2009; Turian et al., 2010; Passos et al., 2014; Agerri and Rigau, 2016). An example of NER is illustrated below.

1) One of the people assassinated in [Sri/B-LOC Lanka/I-LOC] was [Kumar/B-PER Ponnambalam/I-PER].

The sentence contains two named entities: *Sri Lanka* is a location and *Kumar Ponnambalam* is a person. Named entities usually consist of sequences of tokens, e.g., *Sri Lanka*, *European Union*, or *University of Basque Country*, instead of just one word. A robust NER system should identify the correct combinations of words as named entities and classify them into a predefined entity type. Also, a named entity could be linked to various surface forms. For example, *Bill Gates*, *President Gates*, *Mr. Gates*, or *B. Gates* all refer to the same person. Furthermore, the same surface form could indicate different named entities. For instance, the form *Washington* could be mentioned as a person, a location, or as an organization (Khalid et al., 2008), and the form *Europe* could mean a continent, a music band, a magazine, etc (Agerri and Rigau, 2016).

NER systems are used in a wide range of tasks such as named entity disambiguation (Cucerzan, 2007; Han and Sun, 2011; Hoffart et al., 2011; Mendes et al., 2011; Hachey et al., 2013), machine translation (Al-Onaizan and Knight, 2002; Babych and Hartley, 2003; Koehn et al., 2007; Li et al., 2013), coreference resolution (Pradhan et al., 2012), event extraction (Doddington et al., 2004; Ahn, 2006; Ji et al., 2008; Hong et al., 2011; Cybulska and Vossen, 2013; Laparra et al., 2017), sentiment analysis (Liu, 2012; Cambria et al., 2013; Pontiki et al., 2016), and information retrieval (Khalid et al., 2008). Even though many efforts have been devoted to NER, several factors such as corpus size, domain and text genre, and the inherent ambiguity of natural language surface forms still hinders the performance of NER systems.

High performance in NER tasks are achieved by supervised corpus-based approaches, which learn probabilistic models from manually annotated data. However, only a small amount of annotated training data is available for some languages and domains due to its high cost of development. This poses a major obstacle to developing semantic processors whenever there is not manually annotated data for a semantic task in a given language or domain. NER systems are often required to label unseen data with out-of-vocabulary words. If the test set contains words with low frequency or specific to a target domain, it is likely that the performance of the NER system will degrade.

There are more than 6000 languages used in the world. However, nowadays usually only the major languages (English, Spanish, German, Chinese, etc) are supported for NER applications mostly due to the lack of annotated training data. In this context, we aim to take the advantage of rich-resource languages for helping leverage with less-resources. The idea is that as long as annotations are available in one language, that information could be carried over to other languages by exploiting parallel corpora. In order to implement this idea, we propose a cross-lingual approach for named entity recognition via word alignments. We aim to generate named entity tags of a target language automatically from multiple source languages exploiting word alignments via parallel corpora. The assumptions are (1) the word alignment pair should have the same named entity tag inventory across languages; (2) the combination of multiple source languages can improve the quality of the projections.

Figures 1-3 shows an example of cross-lingual projection from German, Spanish, Italian to the same translated English sentence. Figure 4 displays the initial predicted tags for English based on the three source languages. Figure 5 presents the spans of named entity tags for *European Parliament* are corrected since *European* is the beginning of the entity and *Parliament* is inside of the entity. Cross-lingual projection transfers the linguistic features from the three source sentences to the target sentence. The projection paradigm not only helps to alleviate the demand of human effort on corpus annotation, but also it does not require language-specific knowledge or resources (Padó and Lapata, 2009). In addition, it maintains the semantic consistency and word alignment across languages.

To implement the cross-lingual projection, three resources are required: (1) parallel corpora such as Europarl (Koehn, 2005) which offers the translations in a sentence aligned form between all official languages in Europe; (2) a word aligner such as Giza++ (Och and Ney, 2003) to create accurate word alignments from parallel corpora; (3) a semantic tagger such as ixa-pipe-nerc for NER. Most importantly, the annotations obtained from multiple projections are harmonized and possibly, they offer predictions with acceptable quality. We believe that our approach can be applied to generate resources without human intervention for general semantic annotations such as NER, WSD, and SRL for target languages with no manually annotated training data.

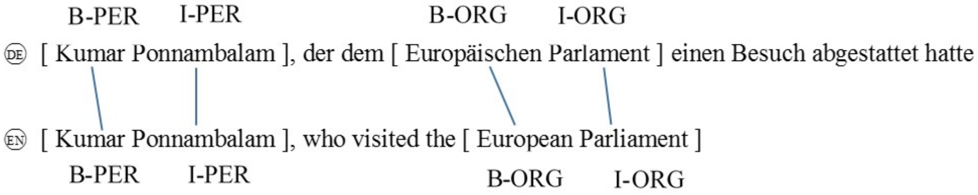

B-PER
I-PER
B-ORG
I-ORG
DE
[Kumar Ponnambalam], der dem [Europäischen Parlament] einen Besuch abgestattet hatte
EN
[Kumar Ponnambalam], who visited the [European Parliament]
B-PER
I-PER
B-ORG
I-ORG

Figure 1: Projecting named entity tags from German to English.

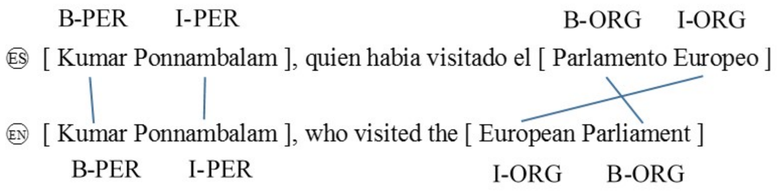

B-PER I-PER B-ORG I-ORG
ES [Kumar Ponnambalam], quien habia visitado el [Parlamento Europeo]
EN [Kumar Ponnambalam], who visited the [European Parliament]
B-PER I-PER I-ORG B-ORG

Figure 2: Projecting named entity tags from Spanish to English.

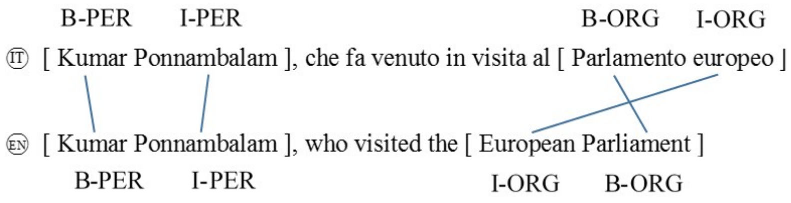

B-PER I-PER B-ORG I-ORG
IT [Kumar Ponnambalam], che fa venuto in visita al [Parlamento europeo]
EN [Kumar Ponnambalam], who visited the [European Parliament]
B-PER I-PER I-ORG B-ORG

Figure 3: Projecting named entity tags from Italian to English.

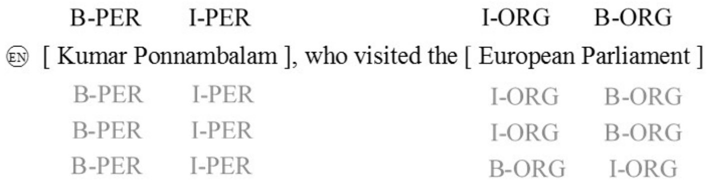


Figure 4: Initial projected named entity tags for English. Tags in black are the final predicted tags and tags in gray are projected tags from ES, IT, DE.

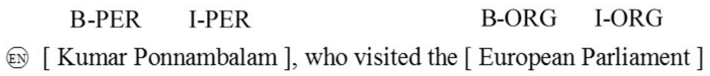

B-PER I-PER B-ORG I-ORG
EN [Kumar Ponnambalam], who visited the [European Parliament]

Figure 5: Final projected named entity tags for English.

1.2 Thesis Outline

The structure of the rest of this thesis is organized as follows. Chapter 2 briefly reviews various approaches of cross-lingual projections, and state-of-the-art in NER. In Chapter

3 we describe the methodology of our research, along with the tools and corpora used. Chapter 4 explains the two approaches of projecting annotations we propose for NER. We then present the results in Chapter 5. Chapter 6 discusses the performances achieved by our NER tagger, and presents an error analysis as well as possible lines to improve our system. Chapter 7 concludes the statement of our study by highlighting the main findings, the contributions, and the possible directions for future work.

2 State-of-the-Art

2.1 Related Work of Named Entity Recognition Classification

The task of named entity recognition was introduced in 1995 by the Sixth Message Understanding Conference (MUC-6) for English language for the uprising need of information extraction from given documents (Grishman and Sundheim, 1996).

The early best performing NER systems that participated in MUC-6 and MUC-7 were rule-based systems which relied on handcrafted rules or lists of gazetteers to detect and classify named entities. The curated rules are manually defined by human experts according to certain patterns for entities using linguistic information such as grammatical, syntactic, and orthographic features. While rule-based approaches can identify complex entities, intensive and costly labor is required for creating rules and the good performance depends on domain-specific knowledge. Furthermore, it may not be adapted to new domains and languages well. Therefore, new approaches which can learn and infer rules from data were introduced including supervised learning, semi-supervised learning, and unsupervised learning.

Furthermore, while MUC 6 was solely devoted to English as target language, the CoNLL shared tasks (2002 and 2003) boosted research on language independent NERC for 3 additional target languages: Dutch, German and Spanish (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). Thus, while in the MUC 6 competition 5 out of 8 systems were rule-based, in CoNLL 2003 16 teams participated in the English task all based on supervised statistical approaches (Nadeau and Sekine, 2007). Supervised learning approaches learn features and induce patterns from labeled data. The best performing systems on CoNLL 2002 and 2003 shared tasks include perceptron (Ratinov and Roth, 2009; Luo et al., 2015; Agerri and Rigau, 2016), and conditional random fields (Passos et al., 2014), among others. Ratinov and Roth (2009) examined the design of NER system in terms of different aspects and developed a system using regularized average perceptron. They compared the representation of text segments, BIO and BILOU where B- represents beginning of, I- represents inside, O- for outside a name entity, and L- for last tokens of multi-token chunks, U- for unit-length chunks. They found out that the BILOU encoding outperformed the BIO encoding on every evaluation. They further built a NER system using non-local features reaching a performance of F1 score 90.5 on CoNLL 2003 shared task. Agerri and Rigau (2016) proposed a new approach of developing multilingual NER system trained with cluster features using averaged perceptron models. The result obtained from their study for the English CoNLL 2003 benchmark is F1 score 91.36 , which is one of the best result reported so far on this dataset.

Recent approaches use neural networks with word embeddings to model the NER dataset. The main reason is to avoid the need of task-specific knowledge and feature engineering. The most commonly applied neural network architectures include convolutional neural networks (CNN) (Chiu and Nichols, 2015; Labeau et al., 2015; Santos and Guimaraes, 2015; Ma and Hovy, 2016), long-short term memory (LSTM) (Hammerton, 2003; Chiu and Nichols, 2015; Ma and Hovy, 2016; Lample et al., 2016), and recurrent neural networks (RNN) (Goller and Kuchler, 1996; Yang et al., 2016). In addition, some researchers combine multiple neural networks together to compensate the disadvantages of each other, obtaining promising results. For example, Collobert et al. (2011) developed a multilayer neural network for various NLP tasks including named entity recognition, part of speech tagging, and semantic role labeling. Chiu and Nichols (2015) followed their study by proposing a new neural network architecture that integrated bidirectional LSTM and CNN architecture to capture word and character information. With this design, no more prior knowledge apart from word embeddings are needed. The reported results on the English CoNLL 2003 benchmark is 90.77 in terms of F1 score.

Lample et al. (2016) proposed a NER system in a model that combining BLSTM and CRF together required no feature engineering or language-specific knowledge. They tested the system on different languages including English, German, Spanish and Dutch. The results showed that for Spanish and German the system outperformed other systems using external resources, and for English and Dutch the system obtained very competitive results.

Ma and Hovy (2016) combined three different neural network architectures for NER and reached good performance. They took the advantages of the characteristics of various neural networks. Firstly, they utilized a CNN to model character-level representations. The representations concatenating with word embeddings are then feed into the BLSTM neural network, which was used as the input to a CRF. In this design, the model learned both the information of a word without requiring task-specific knowledge, and the context sequence around the word. Their experiments tested on CoNLL 2003 shared task reported 91.21 F1 score. Table 1 summarizes the performance of state-of-the-art NER systems using supervised and neural approaches for English. To our knowledge, the NER tagger (Agerri and Rigau, 2016) we use outperforms other state-of-the-art NER systems on several NER benchmarks, including English CoNLL 2003. We can conclude that our results using this tagger will be competitive.

Model	F1
Ratinov and Roth (2009): Perceptron	90.57
Passos et al. (2014): CRF	90.90
Chiu and Nichols (2015): LSTM-CNN *	90.77
Lample et al. (2016): LSTM *	90.94
Luo et al. (2016): Extended Semi-CRF *	91.20
Agerri and Rigau (2016): Perceptron	91.36
Yang et al (2016): RNN *	91.20
Ma and Hovy (2016): LSTM-CNN-CRF *	91.21

Table 1: English NER results on English CoNLL 2003 test set. * indicates the results trained on neural models.

On the other hand, other approaches have been explored to avoid supervision and overcome resource scarcity for NER: (1) silver-standard annotations (Nothman et al., 2013); (2) knowledge from Wikipedia (Toral and Munoz, 2006; Kazama and Torisawa, 2007; Ratinov and Roth, 2009; Nothman et al., 2013); (3) cross-lingual projections (Yarowsky et al., 2001; Xi and Hwa, 2005). Silver-standard annotations are using lower quality but competitive corpora to train supervised systems (Nothman et al., 2013). The general idea of using knowledge from Wikipedia is to exploit the already-available resources to generate annotations for many languages. For instance, Nothman et al. (2013) build multilingual annotations for NER automatically via Wikipedia anchor links. Further descriptions on cross-lingual projection are presented in the following section.

2.2 Approaches to Cross-Lingual Projections

Annotation projection across languages was developed to overcome the problem of resource scarcity (Yarowsky et al., 2001; Xi and Hwa, 2005) in various NLP applications such as machine translation, information retrieval, NER (Yarowsky et al., 2001; Täckström et al., 2012), semantic role labeling (Padó and Lapata, 2009), POS tagging (Yarowsky et al., 2001; Ganchev and Das, 2013; Täckström et al., 2012; Fossum and Abney, 2005), language modeling (Gfeller et al., 2016). Cross-lingual projection refers to utilize aligned pairs of sentences in parallel corpora to obtain the linguistic annotation by taking advantage of the translational equivalences in aligned sentences (Padó and Lapata, 2009).

Previous work on cross-lingual projection focuses mostly on one-to-one projection with one source language as reference to induce linguistic information for a target language. However, we believe that multiple projections would yield precise and robust prediction since semantic annotation should be similar across languages if the text is aligned. Multiple projections were explored only by a few number of studies. Yarowsky et al. (2001) utilized multi-bridging two languages to induce lemmatization for the third language. Fossum and Abney (2005) followed Yarowsky et al. (2001) to train multiple POS taggers from monolingual source data and combine their annotations to project them to a given target language. Therefore, we propose a new approach to apply cross-lingual projection to NER system. An example of cross-lingual projection for NER can be found in 1.1.

Cross-lingual projection faces few problems such as building robust alignments between units of annotation when performing the projection for named entity tags. The aligned unit could be words, constituents, or phrases. When performing the projection it is important to consider the semantic structure shared between the sentences and attain the span of named entity tags (Padó and Lapata, 2009).

Another approach used to leverage the projected annotations is domain adaptation by learning a prediction model for a target domain by exploiting information in a label rich source domain (McClosky, 2010; Carreras and Màrquez, 2005; Daume III and Marcu, 2006; Turian et al., 2010; Faruqui et al., 2010). Information from the rich source can be stored as features or other representations to transfer to other sources. In this sense, Turian et al. (2010) employed word embeddings (Collobert and Weston, 2008) or Brown clusters (Brown et al., 1992) as word features and showed that it can improve out of domain performance of NER systems. Passos et al. (2014) also demonstrated that word2vec embeddings obtained from skip-gram algorithm (Mikolov et al., 2013) can be applied to phrase-based features. Distributed representations are widely exploited in NLP and their effectiveness has been demonstrated for NER (Turian et al., 2010; Agerri and Rigau, 2016) and SRL (Collobert and Weston, 2008), among others. However, more research is required to fully understand which representation offers the best performance for each NLP task, how to effectively combine them, and which unlabeled data is more appropriate for each type of word representation.

A major problem in projection approaches is the translational divergences (Van Leuven-Zwart, 1989; Padó and Lapata, 2009). The idea is that the sentence in the target language might be different from the one in the source language in terms of systematic structures or semantics. The issue might cause misalignments (Padó and Lapata, 2009). For instance, in the study of Yarowsky and Ngai (2001), they successfully utilized word-aligned corpora to create a POS tagger and noun-phrase bracketers, even though the projection annotation was noisy.

3 Methodology

In order to develop our system we need: (i) a named entity recognition 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 for tagging the parallel corpus, and (iv) a gold-standard test set to evaluate our approach. In this chapter we will describe each of the resources used in the rest of our work. For the four languages that are considered in our study: English, Spanish, German, and Italian.

3.1 Named Entity Recognition Tagger

The tool we use for Named Entity Recognition (NER) is ixa-pipe-nerc[1] (Agerri and Rigau, 2016), a multilingual and language-independent tagger included in the IXA pipes tools (Agerri et al., 2014). Ixa-pipe-nerc learns supervised models via the Perceptron algorithm as described by Collins (2002). It is designed to decrease the demand of linguistic motivated features and annotations, such as lemmas, POS tags, and syntax, by exploiting a simple and shallow feature set.

The system consists of: (i) local shallow features, based on orthographic, word shape and n-gram features with their context; (ii) three clustering features, based on unigram matching; (iii) publicly available gazetteers. Specifically, ixa-pipe-nerc implements, on top of the local features, a combination of word representation features: (i) Brown clusters (Brown et al., 1992), taking the 4th, 8th, 12th and 20th node in the path; (ii) Clark clusters (Clark, 2003) and, (iii) Word2Vec clusters (Mikolov et al., 2013), based on K-means applied over the extracted word vectors using the skip-gram algorithm. The implementation of the clustering features looks for the cluster class of the incoming token in one or more of the clustering lexicons induced following the three methods listed above. If found, the class as the feature is added. Brown clusters apply to token related features, which are duplicated.

The ixa-pipe-nerc tagger includes a simple but effective method to *combine* and *stack* various types of clustering features induced over different data sources or corpora, with state of the art results in newswire Named Entity Recognition (Agerri and Rigau, 2016) both for in-domain and out-of-domain evaluations, including the popular CoNLL 2002 and 2003 benchmarks.

3.2 Corpora

Four types of corpora are used:

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.

2) The Europarl parallel corpus on which to perform the cross-lingual projections.

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

3) A Europarl gold-standard test set: a new manually-annotated evaluation set taken from Europarl.

4) Back-off corpora to resolve ties in the projection step.

3.2.1 Gold Out-of-Domain Data for Initial Annotation

As mentioned in section 2.1, The CoNLL 2002 and 2003 shared tasks were proposed to promote language-independent named entity recognition. They covered Dutch, German, English and Spanish and four entity types, namely, LOCATION, MISCELLANEOUS, ORGANIZATION and PERSON. In our work, the CoNLL 2002 shared task (Tjong Kim Sang, 2002) is used for Spanish data and the CoNLL 2003 shared task (Tjong Kim Sang and De Meulder, 2003) for English and German data. CoNLL 2002 consists of the 26 thousand tokens collected from news articles made by the Spanish EFE News Agency for Spanish. CoNLL 2003 contains the 20 thousand tokens from the Reuters Corpus for English and the 20 thousand tokens from the German newspaper Frankfurter Rundshau for German. We decided not to choose Dutch because we wanted to choose two germanic and two romance languages. Thus, the fourth language chosen was Italian. For Italian, we use the Evalita 2009 dataset (Speranza, 2009). It is composed of the 21 thousand tokens from 525 news stories of local newspaper following I-CAB, Italian Content Annotation Bank (Magnini et al., 2007). Evalita 2009 further includes Geo-Political Entity (GPE) while no miscellaneous entity type is annotated.

3.2.2 Europarl

Europarl is a parallel text corpus aligned at sentence level (Koehn, 2005). The corpus is very well-known in the NLP field and it is widely used for statistical machine translation, among other applications. We use Europarl version 7 which contains around 60 million words for each of the 21 languages. Table 2[2] displays the summary of Europarl version 7 for the 4 languages relevant to our work. The sentences are collected from the proceedings of the European Parliament and aligned automatically via the GIZA algorithm introduced by Gale and Church (1993). Europarl will be used to project the NER annotations from three source languages to a fourth target language.

Language (L1-L2)	Sentence Pairs	L1 Words	L2 Words
es-en	1965734	51575748	49093806
de-en	1920209	44548491	47818827
it-en	1909115	47402927	49666692

Table 2: Summary of Europarl version 7. For each language aligned to English, the number of sentence pairs and words (include separated punctuation) from the pair of languages for the sentence aligned corpus (Koehn, 2005) are listed.

[2] Retrieved from: http://www.statmt.org/europarl/index.html

3.2.3 Europarl NER Gold-Standard

In order to be able to evaluate our approach a gold-standard evaluation set is required. Thus, the first 1500 sentences of Europarl were reserved for manual NER annotation. The rest of the Europarl is used as training set.

In order to build the gold standard the following steps were undertaken: The sentences for each pair of languages were word-aligned automatically via Giza++ (Och and Ney, 2003). Then, the sentences without entities in all four languages were discarded. Finally, each word in the remaining 799 sentences was manually annotated at NER level following the CoNLL 2003 annotation guidelines for all 4 languages. Using the CoNLL guidelines would allow us to directly compare the CoNLL 2002 and 2003 trained models with those models that will be generated from the automatically projected data on Europarl. Furthermore, the gold-standard test set will be used in order to build an upper-bound of the algorithm for the projection of annotations.

3.2.4 Back-Off Corpora

Back-off corpora is used to resolve ties when projecting NER annotations. 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. Two corpora are used for the back-off strategy, Wikiner and Europarl:

1) Wikiner[3] is a silver-standard dataset built by categorizing Wikipedia articles into named entity types (Nothman et al., 2013). We choose Wikiner for its large quantities of annotations for 9 languages. We used the Wikiner dataset for back-off using the original named entity annotations of the dataset and an alternative version which consists of annotating again the dataset using ixa-pipe-nerc (best CoNLL 2003 model).

2) The Europarl train set is tagged with ixa-pipe-nerc trained on CoNLL 2003. Table 3 summarizes the number of tokens and entities per type of the Wikiner dataset with the original named entity annotations as this was the back-off dataset that produced best results. While Wikiner was chosen due to its size, Europarl is selected because it is a domain-specific corpora.

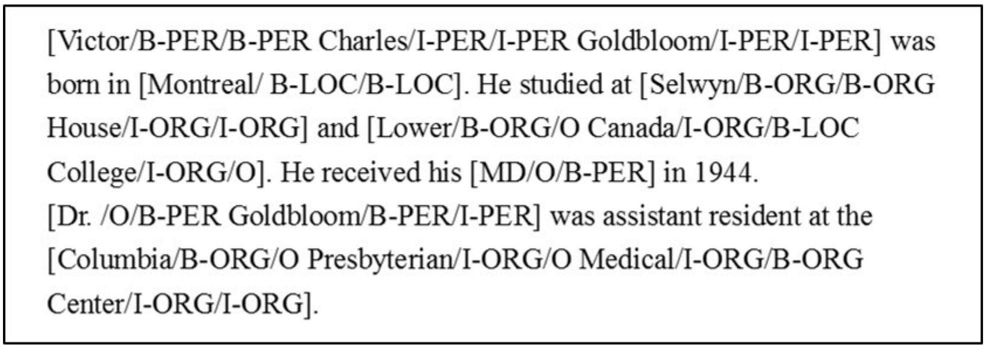
	Articles	Tokens	LOC	PER	ORG
en	3398404	3499655	123210	144600	89446
de	1123266	3499964	272980	215342	134149
it	723722	3499776	200885	114076	34937
es	632400	3500013	182483	116290	41277

Table 3: Wikiner statistics per language and original named entity annotations.

[3] Available at: https://hackage.haskell.org/package/chatter-0.9.1.0/docs/NLP-Corpora-WikiNer.html

3.3 Evaluation Metrics

As it is customary in NER, we use the CoNLL F1 for evaluation. F1 is the harmonic mean of precision and recall. In this metric, only the entity that matches both the class and span of the gold standard entity is a true positive (TP). Any other cases are false positives (FP), false negatives (FN) or both, in case of partial matches. In other words, partial matches are penalized. Figure 6 is an example of annotated text. An example of evaluation in NER is demonstrated as follows. In the text there are 6 gold entities (*Victor Charles Goldbloom*, *Montreal*, *Selwyn House*, *Lower Canada College*, *Goldbloom*, *Columbia Presbyterian Medical Center*) and 7 predicted entities (*Victor Charles Goldbloom*, *Montreal*, *Selwyn House*, *Canada*, *MD*, *Dr.Goldbloom*, *Medical Center*). Only the predicted entities with exact match in the gold standard will be regarded as true positive. That is, there are three true positives in the example, including *Victor Charles Goldbloom*, *Montreal*, and *Selwyn House*. The entities predicted by the system but not in the gold standard, or the ones retrieved correctly but annotated with wrong tags will be regarded as false positives. In the example the 4 false positives are *Canada*, *MD*, *Dr.Goldbloom*, and *Medical Center*. The entities that appear in the gold standard but are not identified by the system will be considered as false negative. In this case, the 3 false negatives are *Lower Canada College*, *Goldbloom*, and *Columbia Presbyterian Medical Center*. According to the F1 definition, precision is the ratio TP / (TP+FP) and recall is the ratio TP / (TP+FN). Therefore, this leads to a precision of 0.43, a recall of 0.5, and a F1 of 0.46.



[Victor/B-PER/B-PER Charles/I-PER/I-PER Goldbloom/I-PER/I-PER] was born in [Montreal/ B-LOC/B-LOC]. He studied at [Selwyn/B-ORG/B-ORG House/I-ORG/I-ORG] and [Lower/B-ORG/O Canada/I-ORG/B-LOC College/I-ORG/O]. He received his [MD/O/B-PER] in 1944.
[Dr. /O/B-PER Goldbloom/B-PER/I-PER] was assistant resident at the [Columbia/B-ORG/O Presbyterian/I-ORG/O Medical/I-ORG/B-ORG Center/I-ORG/I-ORG].

Figure 6: Example of annotated text (Adopted from Atdağ and Labatut (2013)). The notation is token/gold/prediction.

TP: 3 (*Victor Charles Goldbloom*, *Montreal*, *Selwyn House*)

FP: 4 (*Canada*, *MD*, *Dr.Goldbloom*, *Medical Center*)

FN: 3 (*Lower Canada College*, *Goldbloom*, *Columbia Presbyterian Medical Center*)

Precision = TP / (TP+FP) = 3 / (3+4) = 0.43

Recall = TP / (TP+FN) = 3 / (3+3) = 0.5

F1 = 2 * (precision * recall / (precision + recall)) = 0.46

4 Projecting Annotations

As stated in the introduction, our aim is to automatically generate data for training Named Entity Recognition taggers by leveraging parallel corpora. We create a new training corpora automatically for a target language. 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 in the target language obtaining a NER tagger which is fully automatically generated.

The assumptions are (1) word alignment pairs should have the same named entity tag across languages; (2) the combination of multiple sources improves the quality of the projections.

The projection of the named entities annotations using parallel data uses both the automatically obtained named entities and word alignments in the Europarl training set. First, given a word in a sentence of target language, we obtain the aligned words and their named entity class in the three source languages. Next, the named entity tags of the target language are projected based on the candidates collected from the three source languages.

For the first version of our projection system we develop two projection algorithms: (1) strict match projection algorithm for the aim of high precision; (2) upper-bound projection algorithm for the aim of high recall. The strict-match projection algorithm considers at least two agreements among three source languages to determine the final tag for the target language. If that agreement is not reached, we apply a back-off strategy using the named entity tag obtained from computing the most frequent tag for that token in a large automatically annotated corpus (Nothman et al., 2013).

The upper bound system aims at establishing the potential performance of the system, giving us an indication of how well can we project the annotations across languages assuming that the named entity annotations in the source languages are correct. Despite the fact that word alignments have been extensively applied to machine translation, it is crucial to investigate how well word alignments can transport semantic annotations such as named entities. In this sense, using word alignments of manually annotated named entities helps to evaluate the potential of word alignments for projecting multilingual semantic annotations. In the remaining of the chapter we describe the two algorithms.

4.1 Strict Match Projection Algorithm

The aim of the projecting algorithm in this thesis is to project the named entity tags to the target language from three source languages based on word alignments obtained from a parallel corpora. The automatically projected annotations will then be used to train new NER models for the target language.

As an example, take English as target language. The projections come from Spanish, Italian, and German. First, we obtain all word alignments from Spanish, Italian and German to English on the Europarl. Second, we collect the named entity tags for aligned data from Spanish, Italian and German. Third, for all the words in English, we determine the NE tags from projections of Spanish, Italian and German. For each language, the projection algorithm include the following steps:

1) Retrieve alignments: Obtain word alignments from each source language via Europarl.

   Europarl aligned corpora stores which word in one language aligns to which word in another language for each sentence. The first step of the projection algorithm is to retrieve word aligned information of each sentence from aligned corpora. The aligned corpora contains the numbers of lines of sentences. Each line is made up of pairs of aligned numbers, denoting word positions in the sentence. We collect the aligned pairs of word positions from all three source languages. Thus, we learn the aligned word positions from each source language to the target language. For strict match system, we use Europarl.

2) Perform prediction: Determine the prediction of the NE tags of the target language from the annotations in the source languages via projection.

   By now, we have the alignments and named entity tags from the source languages collected in the previous step. The next step in the algorithm is to decide the named entity tags for the target language. It should be noted that different kinds of alignments affect how to generate the projections and the quality of the projections. There are various types of alignments between two languages, 1-1 alignment, multiple alignments, no alignments, and misalignments (See Table 4 for more examples).

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
No alignments	Los; Beschäftigungspakten; NONE	European

Table 4: Examples of various alignments from Spanish, German, Italian to English.

1-1 alignments occurs when to only one NE token (tag) is aligned from the source language to one token (tag) in the target language. Multiple alignments occurs to more than one NE tokens (tags) aligned from the source language to one token (tag). For example, in Table 4, two tokens *del* and *Parlamento* in Spanish are aligned to one token *Parliament* in English. When performing projections, one should develop different strategies for the different alignments.

Generally, we develop different decision rules for predicting NE tags based on two conditions, agreement and span of the named entity tag. The agreement among the source languages should be considered, because the projections in our system come from multiple languages. In order to achieve a competitive system, we should be able to predict and identify the best predicted tag from possible projections. In order to achieve for high precision in the projections, we consider only the candidate tags that are specified by at least two source languages for a given token. If the named entity tag does not agree in at least two source languages, we perform a back-off strategy which is explained in step 4.

3) Assign the span of the named entity tag: The span of the named entity tag is not determined from the alignment. Instead, it is decided in the target language itself. In other words, if the predicted named entity tag is the beginning of that type of named entity tag, it would be B-, and otherwise, it would be I-. Table 5 illustrates how we project the NE tag for the token *European* in English as target language. Take one 3-agreement for instance. First, the projections from the source languages all indicate ORG to English. Hence, the type of entity will be ORG. Second, determine the span of the entity according to the position of the entity in the text. If it is in the beginning of the entity, the final projected tag will be B-ORG. Otherwise, it will be I-ORG.

Alignment	Entity class from es; de; it	Prediction/Projected tag
3-agreement	ORG; ORG; ORG	B-ORG (if it is the beginning of the entity)
2-agreement	ORG; ORG; PER	I-ORG (if it is not the beginning of the entity)
no agreement	ORG; MISC; PER	consult the back-off strategy
more than one 2-agreement	ORG; ORG, LOC; LOC	consult the back-off strategy

Table 5: Example of the projecting algorithm for projecting the token *European* in English as target language.

4) Back-off strategy: Quite often not every word in a source language aligns to another word in another language. A back-off strategy is developed to resolve ties when projecting annotations using the automatically annotated Europarl training data, for those situations where we cannot ensure the named entity tags based on the projections. The situations include more than one three-agreement or two-agreement alignment, or when no agreements are found.

Consider *Paris* as an example of a given token. It could be mentioned as a person, a location, or an organization. If there are no alignments or other contextual indication, it is challenging to recognize to which of them it refers. In this context we believe that leveraging statistical information would be a reasonable solution. The idea is to create a frequent tag dataset which contains the ranks of the possible NE tags for each token based on large corpora. The frequency of occurrence of a tag for a token is an indicator of how likely the token is marked as the tag. Therefore, we exploit the frequent tag dataset to project the most likely NE tag for a given token whenever we cannot perform projection via alignments.

For each language in our experiments, we create the back-off database of named entity tags using three corpora, Wikiner, Wikiner annotated with ixa-pipe-nerc automatically, and Europarl (described in Chapter 3) calculating the frequency of each named entity tag for a given token in the corpora. In the end we obtain a rank of named entity types for a given token in the corpus.

Whenever the back-off strategy is needed, we check if the database contains the token where the back-off strategy is needed. If the database contains the token, the most frequent named entity tag for the token will be the predicted tag. If it is more than one three-agreement or two-agreement alignment and the database does not contain the token, the predicted tag will be the first tag in the alignments since the token and the tag are not available in the database. If there is no alignments (agreements) among the three source languages and the database does not contain the token, we will add *O* as the predicted tag to the token. For instance, in the example of no agreement for the token *European* in Table 5, we consult back-off strategy. We inspect the most frequent named entity tag for the token *European* from the back-off database (See Table 6). According to Table 6, ORG is the most common named entity tag, and therefore ORG will be the predicted tag for the token *European*.

Entity class	Frequency
ORG	2475
PER	943
LOC	537
MISC	31

Table 6: Example of the back-off database for the token *European* in English.

4.2 Upper-Bound Projection Algorithm

The aim of upper-bound projection algorithm is to illustrate how well the system can perform. In order to calculate the upper bound we run the strict match algorithm on the gold standard test data with the condition that all annotations in the three source languages coincide and without the back-off step. The projection across languages is then evaluated on the gold-standard of the target language via the CoNLL script. For each language, the projection algorithm is implemented with the following steps:

1) Retrieve alignments: The same as the first step of strict match projection algorithm, except we obtain word alignments from each source language via gold standard test data.

2) Perform prediction: Determine the prediction of the NE tags of the target language from the annotations in the source languages via projection.

   For upper-bound projection algorithm, we perform prediction in the same way as strict match projection system but focusing on three agreement only. This is to ensure we obtain a prediction which is as precise as possible. Another difference from the strict match projection system is that the prediction is based on three conditions (See Table 7): (a) 1-1 alignment; (b) multiple alignments with the agreed tags across three languages; (c) multiple alignments without the agreed tags across three languages-we add *O* to the token.

Alignments	Tokens	Tags in projection es; de; it	Projected tag
(a) 1-1 alignment	European	ORG; ORG; ORG	ORG
(b) Multiple alignments	Parliament	ORG; O, ORG; ORG	ORG
(c) Multiple alignments	European	ORG; LOC, LOC; ORG	O

Table 7: Examples of performing projection in upper bound system.

3) Assign the span of the named entity tag: The same as the third step of the strict match projection algorithm.

In the next chapter, we will compare these two algorithms on projecting named entity annotations.

5 Empirical Results

In this chapter we first run our upper-bound algorithm described in the previous section to calculate the upper-bound of our projection system. Second, we will apply the strict match algorithm to automatically create training data for every target language. The training data will be used to train new NER models. Finally, we will compare the fully automatically trained with the models trained on out-of-domain, manually annotated gold-standard data for each language (e.g., using the CoNLL and Evalita datasets).

5.1 Upper-Bound Projection

Using strict match various types of possible alignments need to be considered: 1-1 alignment, multiple alignments, and no alignment. 1-1 alignment is the simplest one to decide the prediction. The NE tag of the aligned token corresponds to the predicted NE tag. However, it is more difficult to decide the prediction if multiple alignments or no alignment are found. First, we would like to inspect if the projections of multiple alignments improve the performance. Hence, we display the results projected with 1-1 and multiple alignments projecting on the gold standard data. For 1-1 alignment, we ignore the tokens with multiple alignments and add O to the token.

Table 8 displays the overview of the results of projecting with 1-1 and multiple alignments. The results with multiple alignments are better than those with 1-1 alignment for all four languages. However, the discrepancy between the results with 1-1 and multiple alignments is small for all languages, except for German. The results with 1-1 alignment is 19 points lower than the results with multiple alignments. It shows that multiple alignments provides the crucial information for predicting the NE tags especially for German. Another reason is that, in German, there are many multiple alignments which link to different NE tags. Moreover, in German, compound words are quite common, which might explain why multiple alignments benefit more for German.

	English	German	Italian	Spanish
1-1 alignment	91.47	75.52	91.75	96.32
Multiple alignments	96.01	94.21	93.50	97.34

Table 8: F1 results on upper bound projection.

To fully understand the alignments, we further analyze the results of our system with 1-1 and multiple alignments on different NE types (see Tables 9 and 10). Again, the results with multiple alignments are better than the ones with 1-1 alignment, except for LOCATION class in Italian. For Italian, the result for LOC with 1-1 alignment is slightly better than the one with multiple alignments. For 1-1 alignment, ORG maintains a rather stable performance across languages, above 96 points. For Italian, ORG and PER reach better performance than LOC. For German, ORG and LOC outperform PER. It is worth mentioning that the performances of PER across languages vary the most (from 69.9 for

German to 96.88 for Spanish). Moreover, PER for German performs more than 10 points below other named entity types. Similar patterns can be observed for the projections using multiple alignments.

NER Class	English	German	Italian	Spanish
LOC	92.17	83.02	83.33	93.85
ORG	92.72	85.60	95.99	96.97
PER	84.40	69.90	88.89	96.88

Table 9: F1 results system with 1-1 alignment.

NER Class	English	German	Italian	Spanish
LOC	95.86	93.33	82.49	95.29
ORG	97.72	97.84	96.57	97.52
PER	93.08	87.01	93.17	98.78

Table 10: F1 results with multiple alignments.

Appendix A shows the detailed results with 1-1 alignment and multiple alignments for English, German, Italian, and Spanish.

To sum up, these results demonstrate that our system projects accurately named entity annotations across languages. Hence, we believe applying the strict match algorithm to induce training data for a target language automatically is feasible.

5.2 Full Project Cycle Using Strict Match Algorithm

The general plan is to use strict match to project annotated named entities in the training set automatically, and train NER taggers with the projected annotations (See Figure 7).

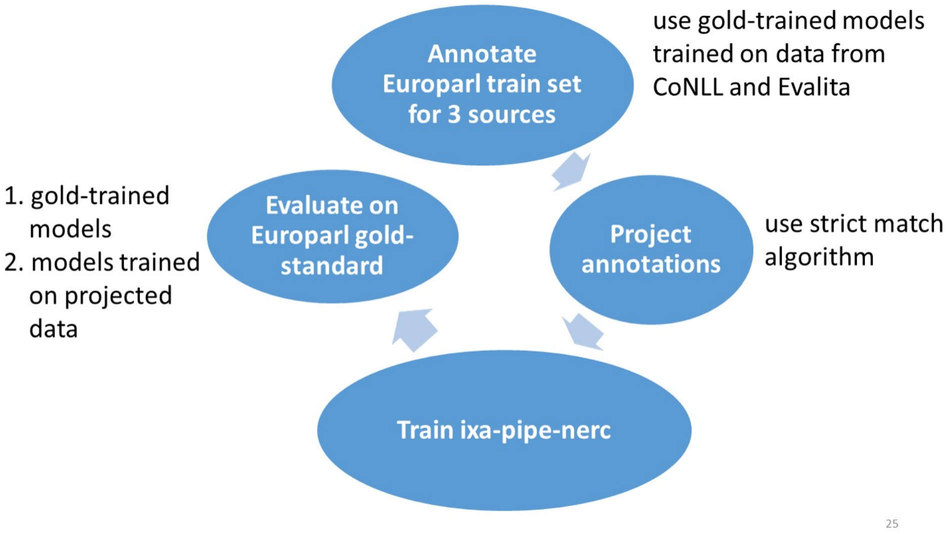

Annotate Europarl train set for 3 sources
use gold-trained models trained on data from CoNLL and Evalita
Project annotations
use strict match algorithm
Train ixa-pipe-nerc
Evaluate on Europarl gold-standard
1. gold-trained models
2. models trained on projected data
25

Figure 7: Full project cycle with strict match algorithm.

In order to do this, for each language the following four steps are undertaken:

1) Annotate Europarl train set using gold-standard models: For each language, we take the Europarl train set and annotate with ixa-pipe-nerc using the best models for each language (plus clusters induced using the Europarl training data) trained on the following gold standard data: CoNLL 2003 for German and English; CoNLL 2002 for Spanish; Evalita 2009 for Italian.

   The best models are the best combinations of features for NERC using ixa-pipe-nerc obtained by Agerri and Rigau (2016). We follow their best models because it is the competitive state-of-the-art compare to other systems. Table 11 lists the detailed settings for the best models with F1 scores for each language:

	Features	F1
en	Brown Reuters + Clark wiki 600 + Word2vec giga 200 + 30 Illinois NER gazetteers (en-91-18)	91.18
de	Clark deWac 500 + Word2vec deWac 100 (de-clusters)	76.42
it	Evalita09 clusters	80.67
es	Brown periodico + Clark giga 400 + Clark wiki 400 + Word2vec giga 400 (es-clusters)	84.16

Table 11: The best features with F1 score in Agerri and Rigau (2016) for annotating Europarl training set.

2) Project annotations using strict match and back-off strategy: We run the strict match projection algorithm using back-off strategy on the annotated training data obtained

from step 1 and Europarl parallel corpus. The projecting algorithm with back-off strategy is described in chapter 4. The projected file is one word with its projected tag, denoting the type of the named entity, per line, following the CoNLL 2002 format.

3) Train ixa-pipe-nerc on the projected data: For each language, we train ixa-pipe-nerc with the projected annotations from step 2.

4) Evaluate both the gold-standard trained models and the models trained on the projected data:

   Hence, for all the languages our models are evaluated on the Europarl gold standard test set to see how well (or bad) did they perform. 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. Furthermore, evaluation of our automatically generated models on the Europarl gold standard will allow us to evaluate our method with respect to the upper bound calculated in section 5.1.

In Table 12 we summarize the dataset used in our work for train and test, obtaining gold standard data and our automatically generated annotation from Europarl.

	Gold-Standard				Europarl			
	LOC	PER	ORG	Tokens	LOC	PER	ORG	Tokens
Train datasets								
en	8286	11128	10001	204567	288037	128561	581770	16822807
de	5229	4495	4241	207484	313781	142623	670524	35620249
it	4013	7001	6313	212478	392593	158831	721913	17920193
es	6804	8224	12382	264715	338979	163255	571351	18127045
Test datasets								
en	1919	2773	2491	46666	145	103	578	22320
de	1284	1822	1262	52098	126	135	464	20187
it	1623	3486	1991	86419	144	100	583	21918
es	1409	1369	2504	51533	157	102	557	23279

Table 12: Datasets used for train and test. Corpora (gold-standard): CoNLL 2003 for German and English; CoNLL 2002 for Spanish; Evalita 2009 for Italian.

5.2.1 Evaluation

Out-of-Domain Evaluation We first perform the out-of-domain evaluation of models on Europarl gold-standard test set. The models using the best combinations of clustering features are applied to the following gold-standard for each language: CoNLL 2002 for Spanish, CoNLL 2003 for English and German, and Evalita 2009 for Italian. The clustering features used for each language are the same best models as described in Table 11. The results are shown in Table 13. With out-of-domain test data, we obtain a F1 score of 64.81 (precision: 70.00, recall: 60.34) for English. The Italian results achieve a F1 score of 64.66

with a precision of 67.03 and a recall of 62.45. For Spanish, a F1 score of 57.60 with a precision of 55.66 and a recall of 59.69 was observed. However, the German results perform much worse than other languages, with a F1 score of 49.87 (precision: 68.40, recall: 39.24).

	Features	Precision	Recall	F1
en	en-91-18-conll03	70.00	60.34	64.81
de	de-clusters-conll03	68.40	39.24	49.87
it	it-clusters-evalita09	67.03	62.45	64.66
es	es-clusters-conll02	55.66	59.69	57.60

Table 13: Evaluating CoNLL and Evalita models on Europarl test.

Among all the languages, German obtains the lowest F1 score, while it achieve comparable performance to English and Italian and higher score than Spanish, in terms of precision. However, the result for recall is not good. The possible reason might be the linguistic characteristic of German. German is characterized by compound words where a word is made up of several words or meanings in other languages. Hence, it is common to see a new word in a corpora and the system fails to identify the named entity. Nevertheless, as long as the system recognizes the named entities, it classifies them to the correct class.

Semi-Supervised Domain Evaluation In addition to out-of-domain evaluation, we also carry out the semi-supervised domain adaptation by adding clustering features induced from the Europarl training set for each language to the same models described in table 11, and based on brown, clark-400 and word2vec-300 clusters for each language. We assess the performance on the Europarl test set.

Similar pattern to out-of-domain evaluation can be found in semi-supervised evaluation, such that English and Italian outperform Spanish and German. The results also show that the performance slightly improves. With semi-supervised-domain evaluation (see Table 14), we obtain a F1 score of 65.08 (precision: 67.49, recall: 62.84) for English. The Italian results achieve a F1 score of 65.82 with a precision of 69.82 and a recall of 62.25. For Spanish, a F1 score of 58.75 with a precision of 58.41 and a recall of 59.10 was observed. However, the German results perform much worse than other languages, with a F1 score of 49.74 (precision: 71.11, recall: 38.25). This is due to the system cannot retrieve the correct named entities (high precision but low recall), compare to other languages.

	Features	Precision	Recall	F1
en	en-clusters-mixed-europarl	67.49	62.84	65.08
de	de-clusters-mixed-europarl	71.11	38.25	49.74
it	it-clusters-mixed-europarl	69.82	62.25	65.82
es	es-clusters-mixed-europarl	58.41	59.10	58.75

Table 14: Evaluating CoNLL and Evalita models with Europarl induced clusters.

Models Trained with Projected Europarl Data We test the models trained with the projected data from Europarl training set on the Europarl gold-standard test set with customized features (labeled as local feature in Table 16). The customized features are generated with local and clustering features based on brown, clark-400 and word2vec-300 clusters for each language (See Table 15).

	Features
en	local: local features clusters: local + brown + clark-400 + word2vec-300 clusters
de	local: local features
it	local: local features w2v-300: local + word2vec-300 clusters
es	local: local features

Table 15: The customized features for models trained with projected Europarl data.

	Features	Precision	Recall	F1
en	en-local-europarl	76.63	60.92	67.88
	en-clusters-europarl	70.30	68.01	69.14
de	de-local-europarl	78.87	63.94	70.62
it	it-local-europarl	71.28	54.82	61.98
	it-w2v-300-europarl	75.14	53.41	62.44
es	es-local-europarl	80.29	53.42	64.16

Table 16: Evaluating models trained on automatically projected data.

The results show that the automatically trained models obtain competitive performances when compared to the gold-standard trained models, especially in terms of precision. In fact, some scores are close to results in standard CoNLL benchmarks, as it can be seen for German and Spanish, where the precision scores are similar to those obtain training and testing on CoNLL data (Agerri and Rigau 2016). These results are very positive and demonstrates that our method can outperform (out-of-domain) gold-standard trained models.

Among the Europarl trained models only the F1 obtained for Italian is inferior to the gold-standard trained model, although this is mostly due to the low recall, given that precision is also substantially higher.

Overall we can say that the results of our first full cycle is very promising and susceptible of being improved in the future, especially looking at ways of increasing the coverage of our models.

On the other hand, we also compare full project cycle and upper bound projection to clearly visualize how the projection on automatic data using strict match compares with the upper bound results. As it can be seen in Table 17, the performance of the models trained with the automatically projected data is still far from the upper-bound established on the gold standard. This means that our method needs to be further improved, especially the strict match algorithm and the way we annotate the training data for initial projections.

	English	German	Italian	Spanish
full project cycle	69.14	70.62	62.44	64.16
upper bound projection	96.01	94.21	93.50	97.34

Table 17: F1 results on comparing full project cycle and upper bound projection.

6 Discussion

Here we first discuss the performance of our models trained with projected Europarl data using strict match projection algorithm, compare to the gold-standard models and upper-bound projection. Second, we present the error analysis on projections. Lastly, we propose several approaches for improving our system.

6.1 Comparison with Gold-Trained Models

In section 5.2 we compare, on our Europarl gold-standard, the CoNLL and Evalita gold-trained models with the automatically induced ones. This evaluation allows to understand if our method produces as good results as the models trained on a gold standard, albeit out-of-domain, data. The F1 results in Table 13 show that the automatically trained models outperform the models trained on gold-standard data except for Italian.

The results also show that our automatically obtained models are particularly good in terms of precision, which means that our strict match projection algorithm is indeed strict. Thus, for English the precision results are 6 points higher, 25 points for Spanish, 10 points for German and 8 points for Italian.

Overall, we believe that this results means that our approach of automatically generating NER taggers is robust and should be further investigated for many more languages.

6.2 Comparison with Upper-Bound Projection

First of all, section 5.1 has shown that our upper bound algorithm applied on gold standard data is highly reliable to transport NER annotations across languages. However, it can be seen that the results obtained using automatic annotations (section 5.2.1) are still quite low compared to the upper bound. This means that although our method performs better than models trained on CoNLL and Evalita data, there is still room for improvement.

6.3 Upper-Bound Error Analysis

Further analysis on type and quantity of errors is performed to better understand the behaviour of our system and to investigate how we can improve it. There are several reasons for performing error analysis. For example, a given token may refer to multiple entity classes, such as *Paris* as a location, a person, or an organization depending on the context of texts. We examine the types of errors, such as whether the errors result from the wrong alignment annotations in the first place (quality of word alignments) or from the projection annotation. Furthermore, we analyze the error types in terms of different entity classes to investigate whether various entity classes behave differently.

Table 18 presents the types of errors with examples for the named entity class PERSON. In the example the word on the left-hand side refers to the token in the target language, and the words on the right-hand side are the aligned token(s) in the source languages separated by a semicolon. The main error made in the projection is due to errors in the automatic word alignment. For instance, the word *date* is aligned to the word *Hicks* which is a name of a person. In addition, the word *reign* is aligned to the word *Forestier* which is a name of a person.

Error type	Word alignments	Multiple alignments between the source languages	Translations
PER	date – Hicks; words - Eieck; was – Vivienne; reign – Forestier, Vivienne; Berenguer – Lieber, verificar, collega, mia	Hicks – 34 jährigen, namens, estgelegt; negli, Stati, Uniti, chiameremo, nome ; llamaremos, con, nombre;	him – Hicks; namens – Hicks; Mr – el, Hicks; him – Hicks
Percentage	77%	14%	9%

Table 18: Examples of errors in named entity class PERSON (token in target language - token in source language).

Other errors come from the existence of multiple alignments between the source languages and due to the translations themselves. In terms of translation, sometimes proper names are replaced in the translation by a coreferential expression. For example, for English the token is *him* (NE class O) which is the translation of the proper name *Hicks* (NE tag PER) in one or more of the source languages. Such translations are the cause of many misalignments in the gold standard data.

Furthermore, *Hicks* is a good example of multiple alignments. The token *Hicks* aligns to 11 tokens from three source languages.

Table 19 presents the examples of the error analysis for the LOCATION named entity class. The types of errors are slightly different to those of the PERSON class. In this case, 46% are span errors. However, this can be due to different forms in different languages. For instance, Table 20 illustrates that the country *the Netherlands* is referred to differently across languages, and so the span changes accordingly. Besides, there are few misalignments. For instance, the token *Sri* in English is misaligned to *Lanka* in Spanish. The token *Lankan* in English is also misaligned to *Sri* in Spanish, and to von in German.

Error type	Word alignments	Span errors	Others* (mismatch between ORG and MISC)
LOC	Sri - Lanka; Lankan – von, Sri; parts – Tauerntunnel, Tauern, Tauri	Parliament; Tunnel; Netherlands; Germany; European; Azores	Europe B-LOC B-MISC; Sri B-MISC B-LOC; Lankan I-MISC I-LOC
Percentage	27%	46%	27%

Table 19: Examples of errors in named entity class LOCATION (*Others is illustrated in this form: token-gold standard-predicted tag).

	Sentence
en	Just think of the road accidents which have occurred over recent years , for example in Belgium , [the/B-LOC Netherlands/I-LOC] and a number of other countries where lorries carrying dangerous goods continued to drive in foggy conditions when really they should have pulled off the road instead .
de	Denken Sie an die Unfälle , die sich in den letzten Jahren im Straßenverkehr ereignet haben . Beispielsweise in Belgien , [den/O] [Niederlanden/B-LOC] und weiteren Ländern fuhren Gefahrguttransporter trotz Nebels weiter und wurden nicht , wie es angesichts der Umstände eigentlich erforderlich gewesen wäre , am Straßenrand abgestellt .
it	Ci sono diversi motivi per dedicare attenzione a questo aspetto , basti guardare quanti incidenti si sono prodotti negli anni scorsi in caso di nebbia in Belgio , o [in/O] [Olanda/B-LOC] o in altri paesi . Tanti sono gli incidenti in cui sono rimasti coinvolti camion che trasportavano merci pericolose e che in tali condizioni meteorologiche non avrebbero dovuto viaggiare invece di trovarsi tranquillamente sulle strade .
es	Fíjense en los accidentes en la carretera de los últimos años . Por ejemplo en Bélgica , [los/O] [Países/B-LOC Bajos/I-LOC] y algunos otros países donde , habiendo niebla , los camiones con mercancías peligrosas seguían en ruta cuando deberían haber parado en esos momentos .

Table 20: Examples of the token (*the Netherlands*) aligned in a sentence across languages.

Finally, Table 21 presents the examples of the error analysis for the ORGANIZATION class. The main errors in ORG come from span error where the system projects the correct named entity class but misses the span. Moreover, some of the errors project ORG as MISC. The possible reason is that the names of organizations vary from language to language and are composed of more than one word. They might use the same word in the other aligned language but in different order.

Error type	Word alignments	Span errors	Others* (mismatch between ORG and MISC)
ORG	concerned - Preussag; supply – Preussag	European; External; Research; Energy; Regional; States;	(token-gold standard-source tag) Community B-ORG B-MISC; European B-ORG B-MISC; Commission I-ORG I-MISC
Total	13%	68%	19%

Table 21: Examples of errors in named entity class ORGANIZATION (*Others is illustrated in this form: token-gold standard-predicted tag).

Summing up, and despite of all the problems discussed, our upper-bound results show that the algorithm we have developed can effectively be used to accurately transport multi word semantic annotations such as named entities via automatic word alignments on parallel data.

6.4 Improvements for the System

Several potential directions could be attempted to improve our system are enumerated below.

1) Refine the projecting algorithm to cover more strategies for various types of alignments. As we mentioned earlier in section 4, there are various types of alignments and it is difficult for our projecting algorithm to cover all the combinations and make precise predictions, especially when multiple alignments exist. For instance, consider the projections *[I-ORG; O, I-LOC; I-ORG, I-LOC]* for a given token, it contains two 2-agreement alignments which are *I-ORG* and *I-LOC*. Without more information it is hard to determine which one is more likely to be the final tag. The possible solutions include to consult the back-off strategy as described in section 4.1, or take other semantic annotations into consideration.

2) Advance the alignments from word level to phrase level. Phrase-based model has been used in machine translation (Och and Ney, 2000; Liang et al., 2006). The intuition of phrase-level alignments is that many-to-many projections can compensate non-compositional phrases.

3) Train the projected annotations with a neural network architecture. There are many lines of research applying neural network to NLP applications and obtaining good results. Our model may be improved by jointly training a LSTM or CNN neural network, which is a good technique for linguistic sequence labeling tasks like NER (Tjong Kim Sang, 2002; Nadeau and Sekine, 2007), to improve the intermediate representations learned in our model.

7 Concluding Remarks

7.1 Summary of the Experiments

A major concern about named entity recognition or any supervised NLP application is the demand of time-consuming and costly hand-crafted annotated corpus. Instead, we propose an approach to transfer semantic annotations to a new language by using already available models and word alignments in other languages. In this way the projected annotations could be used for generating semantic processors for the target language without training data for target language. The assumptions underlying our approach are: First, the word alignment pair entails the same named entity tag across languages. Second, the quality of the projections can be improved by using the combination of multiple sources. This would alleviate the problem of projecting errors as well as offering a clue to examine the various possible strategies for combining the projection mappings.

Based on the results, the two assumptions proposed in 1.1 are confirmed: applying word alignments is viable and effective for inducing semantic processors. Our approach outperforms the gold-trained models except for Italian. Moreover, our models perform consistently well in terms of precision, which demonstrates the robustness of our approach. Additionally, we perform error analysis by examining the NER gold-standard corpora to better understand the behaviour of our projection algorithms.

7.2 Main Contributions

Our contributions are the following:

1) We propose a simple yet robust model for inducing NER taggers automatically using parallel data without any human intervention. To the best of our knowledge, the projecting annotations using word alignment of parallel corpora with multiple languages to induce semantic annotations is novel.

2) Our NER taggers are based on a multilingual architecture, focusing on English, German, Spanish and Italian. Nevertheless, the system is not restricted to these languages as it can be trained and applied to any languages without the demand of language-specific annotations. Furthermore, we are the first few to employ many-to-one projections to advance the quality of projection to the target language.

3) Our automatic generated model achieves similar or even superior results compared to NER taggers trained on manually, out-of-domain, annotated data. Furthermore, our method is designed to meet the requirement of robustness by training our model on ixa-pipe-nerc, a competitive and publicly available state-of-the-art NER.

4) We will release the source code of the system for its practical use in projecting annotations, to guarantee the reproducibility of the results and further applications.

7.3 Future Work

Several potential directions could be attempted to improve our system.

1) Perform another iteration of our whole process again. In the current research, we only implement our method for one iteration (see section 5.1). We believe that if we run a new iteration using the model obtained from the first iteration should improve the performance.

2) Include more languages to improve the quality of projections. In our experiment we consider four languages for projecting annotations (two Romance and two German languages), where three source languages are used to project a named entity tag for a given language. It would be worth to know whether introducing more languages for source annotations will improve the performance, while it might increase the difficulty of performing projections at the same time. Integrating more source annotations is likely to substantially cancel out projection errors. Furthermore, using sources from various language families, such as Chinese or Japanese, may allow to investigate different combinations of mapping strategies and explore if similar results could be obtained.

3) Introduce more semantic annotations. In addition to named entity tags, it is worth discovering different sources to provide deep semantic information about a language, for instance, semantic role labeling (SRL).

4) Carry out out-of-domain evaluations on different datasets in order to access the performance of our system in a real-world situation. Additionally, it is crucial to understand the accomplishment of the system across genres and subject domains. Hence, we will compare the gold-trained models with projected models on out-of-domain data such as the MEANTIME corpus (Minard et al., 2016). If our model achieves similar or better performance to the gold-trained models, we may conclude that our system produces a robust tagger.

References

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

Rodrigo Agerri, Josu Bermudez, and German Rigau. Ixa pipeline: Efficient and ready to use multilingual nlp tools. In *LREC*, volume 2014, pages 3823–3828, 2014.

David Ahn. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8. Association for Computational Linguistics, 2006.

Yaser Al-Onaizan and Kevin Knight. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 400–408. Association for Computational Linguistics, 2002.

Samet Atdağ and Vincent Labatut. A comparison of named entity recognition tools applied to biographical texts. In *Systems and Computer Science (ICSCS), 2013 2nd International Conference on*, pages 228–233. IEEE, 2013.

Bogdan Babych and Anthony Hartley. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*, pages 1–8. Association for Computational Linguistics, 2003.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4): 467–479, 1992.

Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.

Xavier Carreras and Lluís Màrquez. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 152–164. Association for Computational Linguistics, 2005.

Jason PC Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*, 2015.

Alexander Clark. Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 59–66. Association for Computational Linguistics, 2003.

Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics, 2002.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

Silviu Cucerzan. Large-scale named entity disambiguation based on wikipedia data. 2007.

Agata Cybulska and Piek Vossen. Semantic relations between events and their time, locations and participants for event coreference resolution. In *RANLP*, pages 156–163, 2013.

Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, 2006.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, volume 2, pages 837–840, 2004.

Manaal Faruqui, Sebastian Padó, and Maschinelle Sprachverarbeitung. Training and evaluating a german named entity recognizer with semantic generalization. In *KONVENS*, pages 129–133, 2010.

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

William A Gale and Kenneth W Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102, 1993.

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

Beat Gfeller, Vlad Schogol, and Keith Hall. Cross-lingual projection for class-based language models. In *The 54th Annual Meeting of the Association for Computational Linguistics*, page 83, 2016.

Christoph Goller and Andreas Kuchler. Learning task-dependent distributed representations by backpropagation through structure. In *Neural Networks, 1996., IEEE International Conference on*, volume 1, pages 347–352. IEEE, 1996.

Ralph Grishman and Beth Sundheim. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, volume 1, 1996.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R Curran. Evaluating entity linking with wikipedia. *Artificial intelligence*, 194:130–150, 2013.

James Hammerton. Named entity recognition with long short-term memory. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 172–175. Association for Computational Linguistics, 2003.

Xianpei Han and Le Sun. A generative entity-mention model for linking entities with knowledge base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 945–954. Association for Computational Linguistics, 2011.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics, 2011.

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1127–1136. Association for Computational Linguistics, 2011.

Heng Ji, Ralph Grishman, et al. Refining event extraction through cross-document inference. In *ACL*, pages 254–262, 2008.

Jun'ichi Kazama and Kentaro Torisawa. Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 698–707, 2007.

Mahboob Khalid, Valentin Jijkoun, and Maarten De Rijke. The impact of named entity normalization on information retrieval for question answering. *Advances in Information Retrieval*, pages 705–710, 2008.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual* *meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics, 2007.

Matthieu Labeau, Kevin Löser, Alexandre Allauzen, and Rue John von Neumann. Non-lexical neural architecture for fine-grained pos tagging. In *EMNLP*, pages 232–237, 2015.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

Egoitz Laparra, Rodrigo Agerri, Itziar Aldabe, and German Rigau. Multilingual and cross-lingual timeline extraction. *arXiv preprint arXiv:1702.00700*, 2017.

Haibo Li, Jing Zheng, Heng Ji, Qi Li, and Wen Wang. Name-aware machine translation. In *ACL (1)*, pages 604–614, 2013.

Percy Liang, Ben Taskar, and Dan Klein. 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. Association for Computational Linguistics, 2006.

Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. Joint named entity recognition and disambiguation. In *Proc. EMNLP*, pages 879–880, 2015.

Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

Bernardo Magnini, Emanuele Pianta, Manuela Speranza, V Bartalesi Lenzi, and Rachele Sprugnoli. Italian content annotation bank (i-cab): Named entities. Technical report, Technical Report FBK-irst. http://evalita. itc. it/tasks/I-CAB-Report-Named-Entitie s. pdf, 2007.

David McClosky. Any domain parsing: automatic domain adaptation for natural language parsing. 2010.

Pablo N Mendes, Joachim Daiber, Max Jakob, and Christian Bizer. Evaluating dbpedia spotlight for the tac-kbp entity linking task. In *Proceedings of the TACKBP 2011 Workshop*, volume 116, pages 118–120, 2011.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

A-L Minard, Manuela Speranza, Ruben Urizar, Begona Altuna, MGJ van Erp, AM Schoen, CM van Son, et al. Meantime, the newsreader multilingual event and time corpus. 2016.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194: 151–175, 2013.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Franz Josef Och and Hermann Ney. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447. Association for Computational Linguistics, 2000.

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

Alexandre Passos, Vineet Kumar, and Andrew McCallum. Lexicon infused phrase embeddings for named entity resolution. *arXiv preprint arXiv:1404.5367*, 2014.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics, 2016.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics, 2012.

Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics, 2009.

Cicero Nogueira dos Santos and Victor Guimaraes. Boosting named entity recognition with neural character embeddings. *arXiv preprint arXiv:1505.05008*, 2015.

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

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 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, 2012.

Erik F. Tjong Kim Sang. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118853.1118877. URL https://doi.org/10.3115/1118853.1118877.

Erik F. Tjong Kim Sang and Fien De Meulder. 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*, CONLL '03, pages 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1119176.1119195. URL https://doi.org/10.3115/1119176.1119195.

Antonio Toral and Rafael Munoz. A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia. In *Proceedings of EACL*, pages 56–61, 2006.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

Kitty Van Leuven-Zwart. Translation and original: Similarities and dissimilarities, i. *Target. International Journal of Translation Studies*, 1(2):151–181, 1989.

Chenhai Xi and Rebecca Hwa. A backoff model for bootstrapping resources for non-english languages. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 851–858. Association for Computational Linguistics, 2005.

Zhilin Yang, Ruslan Salakhutdinov, and William Cohen. Multi-task cross-lingual sequence tagging from scratch. *arXiv preprint arXiv:1603.06270*, 2016.

David Yarowsky and Grace Ngai. Inducing multilingual pos taggers and np brackets via robust projection across aligned corpora. In *Proceedings of NAACL*, 2001.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 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, 2001.

Appendix A

NER Class	Precision	Recall	F1
LOC	94.64	89.83	92.17
ORG	97.19	88.64	92.72
PER	86.79	82.14	84.40
Total	93.98	89.10	91.47

Table 22: F1 results with 1-1 alignment for English.

NER Class	Precision	Recall	F1
LOC	89.80	77.19	83.02
ORG	95.54	77.54	85.60
PER	83.72	60.00	69.90
Total	88.48	65.87	75.52

Table 23: F1 results with 1-1 alignment for German.

NER Class	Precision	Recall	F1
LOC	81.25	85.53	83.33
ORG	96.14	95.83	95.99
PER	88.89	88.89	88.89
Total	91.75	91.75	91.75

Table 24: F1 results with 1-1 alignment for Italian.

NER Class	Precision	Recall	F1
LOC	93.85	93.85	93.85
ORG	97.26	96.68	96.97
PER	98.41	95.38	96.88
Total	96.89	95.75	96.32

Table 25: F1 results with 1-1 alignment for Spanish.

NER Class	Precision	Recall	F1
LOC	95.29	96.43	95.86
ORG	98.36	97.09	97.72
PER	91.36	94.87	93.08
Total	95.57	96.46	96.01

Table 26: F1 results with multiple alignments for English.

NER Class	Precision	Recall	F1
LOC	90.59	96.25	93.33
ORG	98.01	97.68	97.84
PER	84.81	89.33	87.01
Total	93.04	95.40	94.21

Table 27: F1 results with multiple alignments for German.

NER Class	Precision	Recall	F1
LOC	79.35	85.88	82.49
ORG	95.68	97.48	96.57
PER	91.46	94.94	93.17
Total	92.43	94.59	93.50

Table 28: F1 results with multiple alignments for Italian.

NER Class	Precision	Recall	F1
LOC	94.19	96.43	95.29
ORG	96.92	98.13	97.52
PER	98.78	98.78	98.78
Total	96.93	97.75	97.34

Table 29: F1 results with multiple alignments for Spanish.