Cross-lingual Projections between Languages from Different Families

Mo Yu[1] **Tiejun Zhao**[1] **Yalong Bai**[1] **Hao Tian**[2] **Dianhai Yu**[2]
[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
{yumo,tjzhao,ylbai}@mtlab.hit.edu.cn
[2]Baidu Inc., Beijing, China
{tianhao,yudianhai}@baidu.com

Abstract

Cross-lingual projection methods can benefit from resource-rich languages to improve performances of NLP tasks in resources-scarce languages. However, these methods confronted the difficulty of syntactic differences between languages especially when the pair of languages varies greatly. To make the projection method well-generalize to diverse languages pairs, we enhance the projection method based on word alignments by introducing target-language word representations as features and proposing a novel noise removing method based on these word representations. Experiments showed that our methods improve the performances greatly on projections between English and Chinese.

1 Introduction

Most NLP studies focused on limited languages with large sets of annotated data. English and Chinese are examples of these resource-rich languages. Unfortunately, it is impossible to build sufficient labeled data for all tasks in all languages. To address NLP tasks in resource-scarce languages, cross-lingual projection methods were proposed, which make use of existing resources in resource-rich language (also called ***source language***) to help NLP tasks in resource-scarce language (also named as ***target language***).

There are several types of projection methods. One intuitive and effective method is to build a common feature space for all languages, so that the model trained on one language could be directly used on other languages (McDonald et al., 2011; Täckström et al., 2012). We call it ***direct projection***, which becomes very popular recently. The main limitation of these methods is that target language has to be similar to source language. Otherwise the performance will degrade especially when the orders of phrases between source and target languages differ a lot.

Another common type of projection methods map labels from resource-rich language sentences to resource-scarce ones in a parallel corpus using word alignment information (Yarowsky et al., 2001; Hwa et al., 2005; Das and Petrov, 2011). We refer them as ***projection based on word alignments*** in this paper. Compared to other types of projection methods, this type of methods is more robust to syntactic differences between languages since it trained models on the target side thus following the topology of the target language.

This paper aims to build an accurate projection method with strong generality to various pairs of languages, even when the languages are from different families and are typologically divergent. As far as we know, only a few works focused on this topic (Xia and Lewis 2007; Täckström et al., 2013). We adopted the projection method based on word alignments since it is less affected by language differences. However, such methods also have some disadvantages. Firstly, the models trained on projected data could only cover words and cases appeared in the target side of parallel corpus, making it difficult to generalize to test data in broader domains. Secondly, the performances of these methods are limited by the accuracy of word alignments, especially when words between two languages are not one-one aligned. So the obtained labeled data contains a lot of noises, making the models built on them less accurate.

This paper aims to build an accurate projection method with strong generality to various pairs of languages. We built the method on top of projection method based on word alignments because of its advantage of being less affected by syntactic differences, and proposed two solutions to solve the above two difficulties of this type of methods.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 312–317,
Sofia, Bulgaria, August 4-9 2013.

Firstly, we introduce Brown clusters of target language to make the projection models cover broader cases. Brown clustering is a kind of word representations, which assigns word with similar functions to the same cluster. They can be efficiently learned on large-scale unlabeled data in target language, which is much easier to acquire even when the scales of parallel corpora of minor languages are limited. Brown clusters have been first introduced to the field of cross-lingual projections in (Täckström et al., 2012) and have achieved great improvements on projection between European languages. However, their work was based on the direct projection methods so that it do not work very well between languages from different families as will be shown in Section 3.

Secondly, to reduce the noises in projection, we propose a noise removing method to detect and correct noisy projected labels. The method was also built on Brown clusters, based on the assumption that instances with similar representations of Brown clusters tend to have similar labels. As far as we know, no one has done any research on removing noises based on the space of word representations in the field of NLP.

Using above techniques, we achieved a projection method that adapts well on different language pairs even when the two languages differ enormously. Experiments of NER and POS tagging projection from English to Chinese proved the effectiveness of our methods.

In the rest of our paper, Section 2 describes the proposed cross-lingual projection method. Evaluations are in Section 3. Section 4 gives concluding remarks.

2 Proposed Cross-lingual Projection Methods

In this section, we first briefly introduce the cross-lingual projection method based on word alignments. Then we describe how the word representations (Brown clusters) were used in the projection method. Section 2.3 describes the noise removing methods.

2.1 Projection based on word alignments

In this paper we consider cross-lingual projection based on word alignment, because we want to build projection methods that can be used between language pairs with large differences. Figure 1 shows the procedure of cross-lingual projection methods, taking projection of NER from English to Chinese as an example. Here English is the resource-rich language and Chinese is the target language. First, sentences from the source side of the parallel corpus are labeled by an accurate model in English (e.g., "Rongji Zhu" and "Gan Luo" were labeled as "PER"), since the source language has rich resources to build accurate NER models. Then word alignments are generated from the parallel corpus and serve as a bridge, so that unlabeled words in the target language will get the same labels with words aligning to them in the source language, e.g. the first word '朱(金容)基' in Chinese gets the projected label 'PER', since it is aligned to "Rongji" and "Zhu". In this way, labels in source language sentences are projected to the target sentences.

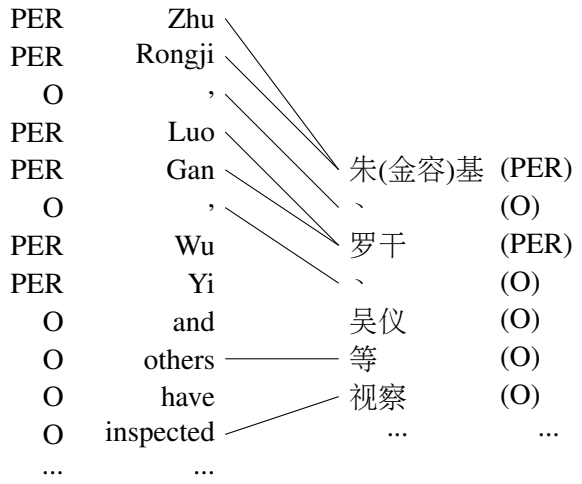

Figure 1: An example of projection of NER. Labels of Chinese sentence (right) in brackets are projected from the source sentence.

From the projection procedure we can see that a labeled dataset of target language is built based on the projected labels from source sentences. The projected dataset has a large size, but with a lot of noises. With this labeled dataset, models of the target language can be trained in a supervised way. Then these models can be used to label sentences in target language. Since the models are trained on the target language, this projection approach is less affected by language differences, comparing with direct projection methods.

2.2 Word Representation features for Cross-lingual Projection

One disadvantage of above method is that the coverage of projected labeled data used for training

Words	$w_{i,i\in\{-2:2\}}, w_{i-1}/w_{i,i\in\{0,1\}}$
Cluster	$c_{i,i\in\{-2:2\}}, c_{i-1}/c_{i,i\in\{-1,2\}}, c_{-1}/c_1$
Transition	$y_{-1}/y_0/\{w_0, c_0, c_{-1}/c_1\}$

Table 1: NER features. c_i is the cluster id of w_i.

target language models are limited by the coverage of parallel corpora. For example in Figure 1, some Chinese politicians in 1990's will be learned as person names, but some names of recent politicians such as "Obama", which did not appeared in the parallel corpus, would not be recognized.

To broader the coverage of the projected data, we introduced word representations as features. Same or similar word representations will be assigned to words appearing in similar contexts, such as person names. Since word representations are trained on large-scale unlabeled sentences in target language, they cover much more words than the parallel corpus does. So the information of a word in projected labeled data will apply to other words with the same or similar representations, even if they did not appear in the parallel data.

In this work we use Brown clusters as word representations on target languages. Brown clustering assigns words to hierarchical clusters according to the distributions of words before and after them. Taking NER as an example, the feature template may contain features shown in Table 1. The cluster id of the word to predict (c_0) and those of context words ($c_i, i \in \{-2, -1, 1, 2\}$), as well as the conjunctions of these clusters were used as features in CRF models in the same way the traditional word features were used. Since Brown clusters are hierarchical, the cluster for each word can be represented as a binary string. So we also use prefix of cluster IDs as features, in order to compensate for clusters containing small number of words. For languages lacking of morphological changes, such as Chinese, there are no pre/suffix or orthography features. However the cluster features are always available for any languages.

2.3 Noise Removing in Word Representation Space

Another disadvantage of the projection method is that the accuracy of projected labels is badly affected by non-literate translation and word alignment errors, making the data contain many noises. For example in Figure 1, the word "吴仪(Wu Yi)" was not labeled as a named entity since it was not aligned to any words in English due to the alignment errors. A more accurate model will be trained if such noises can be reduced.

A direct way to remove the noises is to modify the label of a word to make it consistent with the majority of labels assigned to the same word in the parallel corpus. The method is limited when a word with low frequency has many of its appearances incorrectly labeled because of alignment errors. In this situation the noises are impossible to remove according to the word itself. The error in Figure 1 is an example of this case since the other few occurrences of the word "吴仪(Wu Yi)" also happened to fail to get the correct label.

Such difficulties can be easily solved when we turned to the space of Brown clusters, based on the observation that words in a same cluster tend to have same labels. For example in Figure 1, the word "吴仪(Wu Yi)", "朱(金容)基(Zhu Rongji)" and "罗干(Luo Gan)" are in the same cluster, because they are all names of Chinese politicians and usually appear in similar contexts. Having observed that a large portion of words in this cluster are person names, it is reasonable to modified the label of "吴仪(Wu Yi)" to "PER".

The space of clusters is also less sparse so it is also possible to use combination of the clusters to help noise removing, in order to utilize the context information of data instances. For example, we could represent a instance as bigram of the cluster of target word and that of the previous word. And it is reasonable that its label should be same with other instances with the same cluster bigrams.

The whole noise removing method can be represented as following: Suppose a target word w_i was assigned label y_i during projection with probability of alignment p_i. From the whole projected labeled data, we can get the distribution $p_w(y)$ for the word w_i, the distribution $p_c(y)$ for its cluster c_i and the distribution $p_b(y)$ for the bigram $c_{i-1}c_i$. We choose $y'_i = y'$, which satisfies

$$y' = argmax_y(\delta_{y,y_i}p_i + \Sigma_{x\in\{w,c,b\}}p_x(y)) \quad (1)$$

δ_{y,y_i} is an indicator function, which is 1 when y equals to y_i. In practices, we set $p_{w/c/b}(y)$ to 0 for the ys that make the probability less than 0.5. With the noise removing method, we can build a more accurate labeled dataset based on the projected data and then use it for training models.

3 Experimental Results

3.1 Data Preparation

We took English as resource-rich language and used Chinese to imitate resource-scarce languages, since the two languages differ a lot. We conducted experiments on projections of NER and POS tagging. The resource-scarce languages were assumed to have no training data. For the NER experiments, we used data from People's Daily (April. 1998) as test data (55,177 sentences). The data was converted following the style of Penn Chinese Treebank (CTB) (Xue et al., 2005). For evaluation of projection of POS tagging, we used the test set of CTB. Since English and Chinese have different annotation standards, labels in the two languages were converted to the universal POS tag set (Petrov et al., 2011; Das and Petrov, 2011) so that the labels between the source and target languages were consistent. The universal tag set made the task of POS tagging easier since the fine-grained types are no more cared.

The Brown clusters were trained on Chinese Wikipedia. The bodies of all articles are retained to induce 1000 clusters using the algorithm in (Liang, 2005) . Stanford word segmentor (Tseng et al., 2005) was used for Chinese word segmentation. When English Brown clusters were in need, we trained the word clusters on the tokenized English Wikipedia.

We chose LDC2003E14 as the parallel corpus, which contains about 200,000 sentences. GIZA++ (Och and Ney, 2000) was used to generate word alignments. It is easier to obtain similar amount of parallel sentences between English and minor languages, making the conclusions more general for problems of projection in real applications.

3.2 Performances of NER Projection

Table 2 shows the performances of NER projection. We re-implemented the direct projection method with projected clusters in (Täckström et al., 2012). Although their method was proven to work well on European language pairs, the results showed that projection based on word alignments (WA) worked much better since the source and target languages are from different families.

After we add the clusters trained on Chinese Wikipedia as features as in Section 2.2, a great improvement of about 9 points on the average F1-score of the three entity types was achieved, showing that the word representation features help to recall more named entities in the test set. The performances of all three categories of named entities were improved greatly after adding word representation features. Larger improvements were observed on person names (14.4%). One of the reasons for the improvements is that in Chinese, person names are usually single words. Thus Brown-clustering method can learn good word representations for those entities. Since in test set, most entities that are not covered are person names, Brown clusters helped to increase the recall greatly.

System	avg Prec	avg Rec	avg F1
Direct projection	47.48	28.12	33.91
Proj based on WA	71.6	37.84	47.66
+clusters(from en)	63.96	46.59	53.75
+clusters(ch wiki)	73.44	47.63	**56.60**

Table 2: Performances of NER projection.

In (Täckström et al., 2012), Brown clusters trained on the source side were projected to the target side based on word alignments. Rather than building a same feature space for both the source language and the target language as in (Täckström et al., 2012), we tried to use the projected clusters as features in projection based on word alignments. In this way the two methods used exactly the same resources. In the experiments, we tried to project clusters trained on English Wikipedia to Chinese words. They improved the performance by about 6.1% and the result was about 20% higher than that achieved by the direct projection method, showing that even using exactly the same resources, the proposed method outperformed that in (Täckström et al., 2012) much on diverse language pairs.

Next we studied the effects of noise removing methods. Firstly, we removed noises according to Eq(1), which yielded another huge improvement of about 6% against the best results based on cluster features. Moreover, we conducted experiments to see the effects of each of the three factors. The results show that both the noise removing methods based on words and on clusters achieved improvements between 1.5-2 points. The method based on bigram features got the largest improvement of 3.5 points. It achieved great improvement on person names. This is because a great proportion of the vocabulary was made up of person names, some of which are mixed in clusters with common nouns.

While noise removing method based on clusters failed to recognize them as name entities, cluster bigrams will make use of context information to help the discrimination of these mixed clusters.

System	PER	LOC	ORG	AVG
By Eq(1)	59.77	55.56	72.26	**62.53**
By clusters	49.75	53.10	72.46	58.44
By words	49.00	54.69	70.59	58.09
By bigrams	58.39	55.01	66.88	60.09

Table 3: Performances of noise removing methods

3.3 Performances of POS Projection

In this section we test our method on projection of POS tagging from English to Chinese, to show that our methods can well extend to other NLP tasks. Unlike named entities, POS tags are associated with single words. When one target word is aligned to more than one words with different POS tags on the source side, it is hard to decide which POS tag to choose. So we only retained the data labeled by 1-to-1 alignments, which also contain less noises as pointed out by (Hu et al., 2011). The same feature template as in the experiments of NER was used for training POS taggers.

The results are listed in Table 4. Because of the great differences between English and Chinese, projection based on word alignments worked better than direct projection did. After adding word cluster features and removing noises, an error reduction of 12.7% was achieved.

POS tagging projection can benefit more from our noise removing methods than NER projection could, i.e. noise removing gave rise to a higher improvement (2.7%) than that achieved by adding cluster features on baseline system (1.5%). One possible reason is that our noise removing methods assume that labels are associated with single words, which is more suitable for POS tagging.

Methods	Accuracy
Direct projection (Täckström)	62.71
Projection based on WA	66.68
+clusters (ch wiki)	68.23
+cluster(ch)&noise removing	**70.92**

Table 4: Performances of POS tagging projection.

4 Conclusion and perspectives

In this paper we introduced Brown clusters of target languages to cross-lingual projection and proposed methods for removing noises on projected labels. Experiments showed that both the two techniques could greatly improve the performances and could help the projection method well generalize to languages differ a lot.

Note that although projection methods based on word alignments are less affected by syntactic differences, the topological differences between languages still remain an importance reason for the limitation of performances of cross-lingual projection. In the future we will try to make use of representations of sub-structures to deal with syntactic differences in more complex tasks such as projection of dependency parsing. Future improvements also include combining the direct projection methods based on joint feature representations with the proposed method as well as making use of projected data from multiple languages.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and helpful suggestions. This work was supported by National Natural Science Foundation of China (61173073), and the Key Project of the National High Technology Research and Development Program of China (2011AA01A207).

References

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

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

P.L. Hu, M. Yu, J. Li, C.H. Zhu, and T.J. Zhao. 2011. Semi-supervised learning framework for cross-lingual projection. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, volume 3, pages 213–216. IEEE.

R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering*, 11(3):311–326.

W. Jiang and Q. Liu. 2010. Dependency parsing and projection based on word-pair classification. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL*, volume 10, pages 12–20.

P. Liang. 2005. *Semi-supervised learning for natural language*. Ph.D. thesis, Massachusetts Institute of Technology.

R. McDonald, S. Petrov, and K. Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 62–72. Association for Computational Linguistics.

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

S. Petrov, D. Das, and R. McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

O. Täckström, R. McDonald, and J. Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure.

O Täckström, R McDonald, and J Nivre. 2013. Target language adaptation of discriminative transfer parsers. *Proceedings of NAACL-HLT*.

H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, volume 171. Jeju Island, Korea.

F Xia and W Lewis. 2007. Multilingual structural projection across interlinear text. In *Proc. of the Conference on Human Language Technologies (HLT/NAACL 2007)*, pages 452–459.

N. Xue, F. Xia, F.D. Chiou, and M. Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207.

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