Journal of Computational Information Systems 9: 6 (2013) 2193–2200
Available at http://www.Jofcis.com

A Cross-lingual Approach for Opinion Holder Extraction [⋆]

Lin GUI, Ruifeng XU[*], Jun XU, Chenxiang LIU

Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China

Abstract

Opinion holder extraction is an important subtask in opinion analysis. However, the lack of fine-annotated corpus influences the supervised machine learning based approach. The cross-lingual approach which leverages the opinion annotation information on other languages is considered as a feasible solution. In this paper, we propose a new cross-lingual opinion analysis approach to leverage fine-annotated opinion corpus. We firstly generate the translations and corresponding annotations of MPQA, which is most important opinion corpus in English, through cross-lingual projection. The obtained transferred corpus is used as the supplementary training data to train a classifier based on Tree Kennel Support Vector Machine (TK-SVM). The experiments on NTCIR-7 MOAT dataset (Simplified Chinese side) show that the proposed cross-lingual approach achieves a higher performance compared to CRFs models.

Keywords: Cross-lingual Opinion Analysis; Opinion Holder Extraction; Tree Kernel Support Vector Machine

1 Introduction

With the development of social network such as Twitter or Micro-blog in recent years, opinion analysis which identifies the subjective expressions and determines their polarity has attracted growing interests. Generally speaking, the existing opinion analysis are camped into coarse grained classification which classifies the sentence or documents into subjective/objective and positive/negative/neutral, and fined-grained information extraction which extracts the opinion components from the opinionated sentences. The latter one is essential to many practical applications.

The recent fined-grained information extraction technique may be roughly camped into rule-/pattern- based approach and machine learning based approach. The former approach requires manual intervention. It achieves good performance in some specific task. However, applying this approach to new tasks is hard [1]. The latter approach is based on some supervised machine learning algorithms such as Conditional Random Fields (CRFs) and Markov model [2,3]. The success of this approach requires large training data which is conflict to the rare fined-annotated

[⋆] http://english.51ielts.com/a/esp/office/2010/1029/30266.html

[*] Corresponding author.

Email address: xuruifeng.hitsz@gmail.com (Ruifeng XU).

1553–9105 / Copyright © 2013 Binary Information Press
March 15, 2013

corpus. This motivated the idea of cross-lingual opinion analysis that transfer fined-grained annotation corpus in other language (define as **Source Language**, SL) to the **Target Language** (TL) as the supplementary training data. The present cross-lingual opinion analysis methods mainly focus on sentence level or document level transfer. The key point is to eliminate translation bias in cross-lingual process. In the fined-grained condition, the machine translation system brings much mis-order bias while the existing fined-grained opinion analysis techniques, especially the ones based on sequence labeling, are sensitive to word order. This puzzles the application of cross-lingual fined-grained analysis.

In this study, we propose a new cross-lingual framework for fined-grained opinion analysis, in particular, for opinion holder extraction. Firstly, we transfer the MPQA, which is the most important opinion corpus in English, to Chinese through machine translation and cross-lingual annotation projection. Considering that the sub-structures in parse-tree are more stable than whole sentences during the cross-lingual projection, we propose to utilize the sub-structure in the transferred opinion corpus to train a classifier based on Tree Kernel Support Vector Machines for identifying opinion holders in Chinese text. The proposed cross-lingual approach is evaluated on NTCIR-7 MOAT Simplified Chinese dataset. The achieved performance is better than the CRFs model with monolingual training data which shows the effectiveness of cross-lingual opinion analysis.

The rest of this paper is organized as follows: Section 2 reviews the related work. The proposed cross-lingual fined-grained opinion analysis approach is presented in Section 3. Section 4 gives the experimental results and discussions. Finally, Section 5 concludes this paper.

2 Related Work

Generally speaking, most existing techniques for fined-grained opinion analysis may be camped into rule-/pattern-based approach and machine learning based approach. Xu et al. proposed to use 62 manual complied rules and patterns are applied to detect the center and seek the boundary of each possible holder and target [1]. This method achieved good results in NTCIR-07 holder & target extraction task. However, these rules and patterns are compiled for specific task and thus it is hard to applying to other tasks.

There are many works on opinion holder/target extraction following the machine learning based approach, especially the sequence labeling models. The typical works including the method based on Markov model [2] and the method based on CRFs model [3]. Furthermore, Moschitti applied convolution kernel on parse-tree, which is proposed by Collins [4] to recognize the argument in English sentences [5]. M. Wiegand [6] proposed to use tree kernel SVM to extract holder from opinion sentence. Normally, the tree kernel based techniques is more efficient compared to the sequence labeling models. It is also robust to the order of sequences. The shortcoming of machine learning based approach is the requirement of large training data which is not available.

Current cross lingual opinion analysis researches focus on the good use of transferred data. J. Xu proposed transfer Ada-Boosting model to selecting instance-level transferred data [7]. Lu proposed a cross lingual method based on EM algorithm for sentence level opinion analysis [8]. Kim proposed a fine-grained cross-lingual opinion analysis method based on parallel corpus [9].

3 Cross-lingual Opinion Holder Extraction based on Tree Kernel SVM

The observations show that both word translation errors and mis-order bias widely existing in the machine translation results. These problems, especially the mi-order bias, puzzled the cross-lingual fine-grained analysis based on sequence labeling models. Considering that the sub-structures in parse tree are more stable than the original sentence during the cross-lingual projection and the Tree Kernel SVM is shown robust to the order of sequences, in this study, we propose a new cross-lingual opinion holder extraction approach based on Tree Kernel SVM (TK-SVM). We firstly transfer the MPQA, an English fined annotated corpus, to Chinese through machine translation and cross-lingual annotation projection. By selecting the sub-structures in parse tree as the main discriminative feature, the transferred opinionated sentences are used as the supplementary samples for training a classifier based on Tree Kernel SVM. The classifier is then used to identify opinion holder in the testing opinionated sentences.

3.1 Translation of MPQA

In this study, we use MPQA as the source language resource and NTCIR-07 simplified Chinese training samples as the target language samples. MPQA contains five different fined-grained labels in which we only use three of them: opinion holder, opinion target and opinion predict words. Considering that NTCIR-07 dataset only provides the annotations of opinion holder and opinion target, 1,410 opinionated sentences are selected from MPQA which contains an annotated opinion predict word and only one annotated holder or target.

We apply BAIDU machine translation (MT) system to translate selected sentences in MPQA to Chinese. This MT system is a phrase based one. We then utilize the provided phrase information to re-construct the sub-structures of annotations in target language. The MT results contain two parts. The main information part is the translation of input sentence. The alignment part gives the aligning information at phrase level. We take an opinionated sentence in MPQA as an instance (shown in Table 1). In this instance, the annotated opinion holder is “Russia”, opinion predict word is “favours” and opinion target is “creation of international instruments”. In this case, the boundaries of all annotations match the translation alignment perfectly. But in all of 1,410 sentences, only 317 ones are matched perfectly like this case. For the unmatched samples, we use GIZA++ to generate word alignment information. Based on this, the phrase annotations may be aligned through a post processing.

Table 1: BAIDU MT result of an opinionated sentence

Source sentence	Russia favours creation of “international instruments” to regulate emissions
Main information	*俄罗斯赞成建立“国际文书”调节排量*
Alignment information	[0, “俄罗斯”, [“0—6”], [Russia], [“0—6”]] [1, “赞成”, [“7—7”], [favours], [“7—7”]] [2, “建立 “”, [”15—13”], [creation of ”], [“15—13”]] [3, “国际文书”, [“28—25”], [international instruments], [“28—25”]] [4, “” ”, [“53—1”], [”], [“53—1”]] [5, “调节排量”, [“55—21”], [to regulate emissions], [“55—21”]]

3.2 Sample construction and cross-lingual projection

3.2.1 Sub-structure construction

In this study, we focus on the identification of opinion holder and opinion target. The training dataset of NTCIR-7(Simplified Chinese side) consists of 428 sentences in which 78 ones have the opinion holder or/and opinion target annotations. Furthermore, we define the verb expressing an opinion from the holder to the target as opinion predict word (OPW). We manually annotated the OPW in these 78 sentences. Next, we use the Standford parser to generate the parse tree of these sentences. Corresponding to the annotated opinion holder, the minimum sub tree whose nodes covers all words in the holder is defined as the holder sub tree. Similarly, the target sub tree is defined. The combination of holder sub tree, OPW, and the shortest path between them in the parse-tree constructs a holder-predicate sub-structure. This sub-structure maintains the syntactic restriction between opinion holder and OPW. Under the similar definition, we obtain the target-predict sub-structures. The examples of holder-predict and target-predict sub-structures in a Chinese opinionated sentence are shown in Fig. 1.

For the TK-SVM classifier, the training samples are these sub-structures. The 110 holder-predict and target-predict sub-structures in these 78 sentences (in the target language) are labeled as Target-language Training Samples, in short TTS.

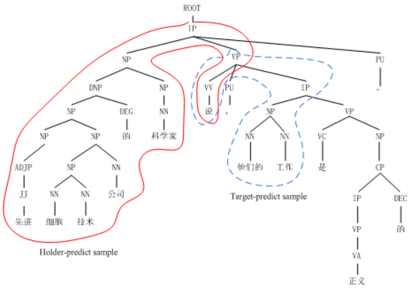

Fig. 1: Sample of Holder-Predict and Target-Predict sub-structures

3.2.2 Cross-lingual sub-structures extraction

In cross-lingual condition, we obviously cannot directly extract the holder-predict and target-predict sub-structures from the parsing tree of translated SL samples with good accuracy because the existence of translation errors lead mass of noisy in the automatic parsing result. Thus, we propose a four step method to extract correct sub-structures from translated SL samples with illustrations:

1) Generate the parse tree for samples on source language (labeled SLS). It is expected to achieve a good parsing accuracy. For a given sentence in SL, *Britain's Prescott pushes for Kyoto Pact Backing*, its parse tree and corresponding sub trees is shown in Fig. 2.

2) Align the opinion holder, OPW and opinion target between original SLS and translated SLS as shown in Table 2.

3) Parse the translated opinion holder, OPW and opinion target, respectively. Since the structures of these words/phrases are relatively simple with short distance, the parsing accuracy for

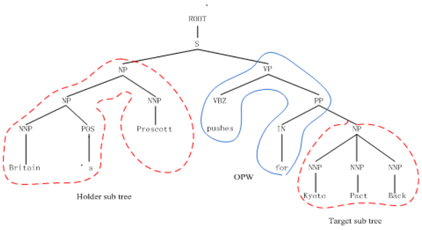

Fig. 2: Parse-tree in source language

Table 2: BAIDU MT result of an opinionated sentence

Source sentence	Britain's Prescott pushes for Kyoto Pact Backing
Translation	*英国的普雷斯科特推动京都协议的支持*
Annotation	Holder: Britain's Prescott(英国的普雷斯科特) OPW: pushes for(推动) Target: Kyoto Pact Backing(京都协定的支持)

these phrases may achieve a good accuracy. Now, the holder/ target sub trees on SL and TL are obtained, respectively, as shown in Fig. 3.

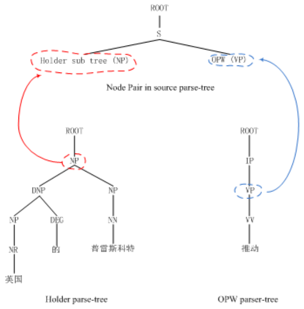

Fig. 3: Cross-lingual sub-tree alignment

4) Replace the ROOT of holder/target sub trees on TL by using the label of holder/target sub trees in the SL parse tree. Meanwhile, the node information in the SL parse tree is assigned to the corresponding node on TL. In this case, the ROOT of holder on TL is replaced by NP and the node information in SL parse tree is assigned to this node.

In this way, even though we did not generate the complete parse tree on TL, we obtain the holder/target sub-structures with good accuracy.

3.2.3 Cross-lingual node tags correction

The Stanford parser adopted different tag set for English and simplified Chinese. Thus, the nod tags correction is required during sub tree projection from English to simplified Chinese. The corrections may be classified as two categories: necessary correction (2 types) and modified correction (4 types).

Necessary correction means this correction must be taken. For example, the simple sentence in English is tagged as “S”, while in Chinese the corresponding tag is “IP”. So, the tags for all simple sentences on TL are corrected as "IP". The training samples with necessary correction are named as original samples (OS).

Modified correction aims to fix the bias in translation and parser system. For example, Stanford parser tends to parse one-word Chinese phrase as noun even if this word may be both noun and verb in semantic level. Thus, if we found a one-word Chinese phrase is tagged as noun but it appears in the OPW, its tag is corrected as verb. The training samples generate by modified correction are named as fixed samples (FS). Through the translation, alignment and projection, there are 1,782 OS and FS samples on TL are obtained. They are combined with TTS (110 samples) to train a TK-SVM classifier.

3.3 TK-SVM classifier

Here, the SVMlight-TK-1.2.1 is utilized to construct the TK-SVM classifier. The classifier is trained by using the mentioned samples. It is noted here that, we do not construct a classifier to determine whether a noun phrase is opinion holder or opinion target. We just utilize the output value of TK-SVM to rank the noun phrases according to the “likeness” to be an opinion holder. If the most like noun phrase with a positive output value in TK-SVM, it is regarded as an opinion holder. If this phrase has a negative value, we consider this sentence has no intuitional holder.

4 Evaluation and Discussion

The proposed cross-lingual opinion extraction approach is evaluated on NTCIR-7 MOAT (simplified Chinese side) opinion holder extraction dataset. The NTCIR-7 test samples contain 1869 opinionated sentences in which 1159 ones has explicit holder. The other 710 sentences have hidden or contextual holders, i.e. the holders are author of the article or co-reference element in other sentences. Our experiment is performed on the whole test set of 1869 sentences. The traditional metrics proposed by NTCIR is adopted [10]. There is one unreasonable point that in this metric that if a system proposes few results with high precision, the P/R/F measure will be fairly high. So, in our experiment the number of correctly extracted holders is also evaluated.

In the first experiment, the proposed TK-SVM based classifier is trained by using different set of samples, respectively. The achieved performances are given in Table 3. It is observed that the classifier trained by TTS achieved the better performance while the individual use of cross-lingual transferred samples leads to lower performance. It attributes to the use of noisy cross-lingual samples. It is also observed that the node tags correction filter out some noises. In addition, the use of FS achieves a better performance compared with OS.

Table 3: Performances of the TK-SVM classifier with different training samples

Training samples	P/R/F	#Correct holder extract
TTS	**0.4586**	**632**
OS	0.3945	533
FS	0.4157	560

In the second experiment, we add samples from OS or FS to TTS randomly, and train language model on the combined samples to observe the variation of performance, which is shown in Fig.

4. It is observed that in most cases the performance improves with the addition of cross-lingual samples in the beginning. However, with the use of more cross-lingual transferred samples, the performance increase slower even decreases. It shows that the supplementary cross-lingual transferred samples lead to the performance improvement at the beginning. However, if more transferred samples are used, the noisy in them start to influence the performance of language model.

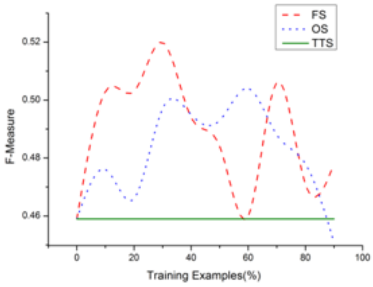

Fig. 4: Performances of the TK-SVM classifier with increasing training samples

The performance achieved by our approach is compared with several reference systems in NTCIR-7 evaluation. To handle the hidden and contextual authors in NTCIR-7 test set, the corresponding module developed by WIA group is adopted while our approach is applied to handle the explicit holder recognition. The performances comparison is given in Table 4. In NTCIR-7 MOAT evaluation, ICLPKU and NLPR systems are both based on CRFs model while WIA uses a set of manually compiled rules and templates (TTRD and NTU did not provide the description their systems). The performance achieved by ICLPKU and NLPR are similar to our TK-SVM classifier with only TTS. With the use of cross-lingual transferred samples, our classifier achieves much better performance. It shows the contribution of cross lingual transfer of fined grained annotated samples. WIA achieves the highest performance, but this method highly depends on manual templates compile and thus the expandability of this method is not satisfactory.

Table 4: Performance comparison with the reference systems

Team	P/R/F	#Correct holder extract
ICLPKU	0.4124	617
TTRD	0.127	106
NLPR	0.4497	407
NTU	0.2909	331
WIA	0.6656	1021
TK-SVM (TTS+OS)	0.554	701

5 Conclusions

In this paper, we proposed a cross lingual opinion analysis framework for opinion holder extraction. The samples in MPQA corpus with fined grained annotations are translated and projected to simplified Chinese as the supplementary samples for training a classifier based on TK-SVM. The evaluations on NTCIR-7 opinion dataset (Simplified Chinese side) show that our approach achieves an encouraging performance which shows the effectiveness of cross lingual method in fine-grained opinion analysis.

Acknowledgement

This research is supported by HIT.NSFIR.201012 from Harbin Institute of Technology, China, Shenzhen Fundamental Research funding No.JCYJ20120613152557576 and Natural Science Foundation of China, No.61203378.

References

[1] R. Xu and K. F. Wong. Coarse-Fine Opinion Mining - WIA in NTCIR-7 MOAT Task. In *Proceedings of NTCIR-7 Workshop Meeting*, Decebmer 16 – 19, 2008.

[2] Y. Choi, E. Breck and C. Cardie. Joint Extraction of Entities and Relations for Opinion Recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 431 – 439, 2006.

[3] Y. Choi, C. Cardir, E. Riloff and S. Patwardhan. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 355 – 362, Vancouver, October 2005.

[4] M. Collins and N. Duffy. New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 263 – 270.

[5] A. Moschitti. A study on Convolution Kernels for Shallow Semantic Parsing. In *proceedings of the 42-th Conference on Association for Computational Linguistic (ACL-2004)*, Barcelona, Spain, 2004.

[6] M. Wiegand and D. Klakow. Convolution Kernels for Opinion Holder Extraction. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 795 – 803, 2010.

[7] J. Xu, R. Xu, Y. Ding and X. Wang. Cross Lingual Opinion Analysis via Transfer Learning. In *Australian Journal of Intelligent Information Processing Systems, Vol 11, No 2 (2010): Computational Neuroscience and Cognitive Science.*

[8] B. Lu, C. Tan C. Cardie and B. K. Tsou. Joint Bilingual Sentiment Classification with Unlabeled Parallel Corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 320 – 330, 2011.

[9] S. Kim, M. Jeong, J. Lee and G. G. Lee. A Cross-lingual Annotation Projection Approach for Relation Detection. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 564 – 571, 2010.

[10] Y. Seki, D. K. Evans, L, -W, Ku, L. Sun, H, -H, Chen and N. Kando. Overview of Multilingual Opinion Analysis Task at NTCIR-7. In *Proceedings of NTCIR-7 Workshop Meeting*, Decebmer 16 – 19, 2008.