Lost in Translations? Building Sentiment Lexicons Using Context Based Machine Translation

Xinfan Meng[1] *Furu Wei*[2] *Ge Xu*[1,3] *Longkai Zhang*[1]
Xiaohua Liu[2] *Ming Zhou*[2] *Houfeng Wang*[1*]
(1) MOE Key Lab of Computational Linguistics, Peking University
(2) Microsoft Research Asia
(3) Department of Computer Science, Minjiang University
mxf@pku.edu.cn, xuge@pku.edu.cn, fuwei@microsoft.com, zhlongk@qq.com,
xiaoliu@microsoft.com, mingzhou@microsoft.com, wanghf@pku.edu.cn

ABSTRACT
In this paper, we propose a simple yet effective approach to automatically building sentiment lexicons from English sentiment lexicons using publicly available online machine translation services. The method does not rely on any semantic resources or bilingual dictionaries, and can be applied to many languages. We propose to overcome the low coverage problem through putting each English sentiment word into different contexts to generate different phrases, which effectively prompts the machine translation engine to return different translations for the same English sentiment word. Experiment results on building a Chinese sentiment lexicon (available at https://github.com/fannix/Chinese-Sentiment-Lexicon) show that the proposed approach significantly improves the coverage of the sentiment lexicon while achieving relatively high precision.

KEYWORDS: Sentiment analysis, Multilingual, Dictionary.

*corresponding author

Proceedings of COLING 2012: Posters, pages 829–838,
COLING 2012, Mumbai, December 2012.

1 Introduction and Related Work

Sentiment lexicons are valuable resources for sentiment analysis; they can be used to identify sentiment words and expression, and they can also be used to generate informative features for sentiment classification of documents. Several sentiment lexicons have been compiled (Stone et al., 1966; Hu and Liu, 2004; Wilson et al., 2005) for English. They are widely used in the research on sentiment analysis. By contrast, due to the high cost of manually compiling a lexicon, sentiment lexicons in many other languages are very few or even unavailable. The shortage of sentiment lexicons limits our capability to analyze the sentiment conveyed in the documents written in other languages; it is estimated that as of May 31 2011, only 26.8% of Internet users speak English [1].

There is some research on automatically building sentiment lexicons for other languages using translation based methods or bootstrapping methods. Straightforward translation methods make use of multilingual dictionaries, and bootstrapping methods enlarge the sentiment lexicons from English sentiment seed words using semantic resources. However, straightforward translation methods suffer from low sentiment word coverage in the bilingual dictionaries. Moreover, in many cases, two or more English sentiment words often are translated to the same foreign word. Both factors lead to smaller translated sentiment lexicons than the original ones. (Mihalcea et al., 2007) study the effectiveness of translating English sentiment lexicon to Romanian using two bilingual dictionaries. The original English sentiment lexicon contains 6,856 entries; after translation, only 4,983 entries are left in the Romanian sentiment lexicon. About 2000 entries are lost or conflated into other entries during the translation process. The translation method is also used in (Wan, 2008, 2011).

On the other hand, though bootstrapping methods don't use bilingual dictionaries and hence are not subject to the limitation of the translation methods, they have relatively high demands for semantic resources such as WordNet (Fellbaum, 1998). Bootstrapping methods enlarge the sentiment lexicons from English sentiment seed words. (Hassan et al., 2011) present a method to identify the sentiment polarity of foreign words by using WordNet (or similar semantic resources) in the target foreign language. (Ku and Chen, 2007) create a Chinese Lexicon by translating the General Inquirer, combining with Chinese Network Sentiment Dictionary, and conducting expansion using two thesauri. Other semi-supervised lexicon construction methods such as random walk (Esuli and Sebastiani, 2006), label propagation (Rao and Ravichandran, 2009; Xu et al., 2010) or graph propagation (Kerry and McDonald, 2010) can also be used here. However, all those methods require high quality lexicon seed words in the target languages and/or some semantic resources, which are not always available in the target languages.

Besides automatic methods, semi-automatic approaches are also studied. (Steinberger et al., 2012) first produce high-level gold-standard sentiment dictionaries for two languages, then translate them automatically into third languages respectively and obtain overlap of translated lexicon. The experiment suggests that this triangulation method works significantly better than simple translation method. However, in some intermediate stages, the dictionaries need to be filtered and expanded manually.

In this paper, we present a simple yet effective approach to creating high quality sentiment lexicons using English sentiment lexicons. Instead of relying on bilingual dictionaries or

[1] http://www.internetworldstats.com/stats7.htm

Context	English	Translation
None	elegant	优雅
	graceful	优雅
Collocation	graceful voice	优美的声音
	graceful dance	曼妙的舞姿
Coordinated phrase	elegant and graceful	典雅大方
	graceful and elegant	雍容典雅
Punctuation	graceful.	优美。
	elegant.	优雅。
	graceful and elegant.	婉约和优雅。

Table 1: Chinese Translations of "graceful" and "elegant" in different contexts

semantic resources, we leverage online machine translation services, which are readily accessible. In order to overcome the word coverage problem, we put each English sentiment word in different contexts to generate different phrases, which can prompt translation engines to return different translations for the same English sentiment word. In particular, we develop three techniques for constructing contexts and generating different phrases. It should also be emphasized that leveraging online machine translation service enables us to easily construct lexicons in many languages; as an empirical study, we use this approach to construct a Chinese sentiment lexicon, and the obtained lexicon is both large and accurate.

2 Our Approach

Formally, our task is to build a sentiment lexicon for a target language, such as Chinese, given an English sentiment lexicon. We use Table 1 to illustrate the idea. As an example, we translate two English positive words, "graceful" and "elegant", to Chinese. When we translate "graceful" or "elegant" individually, they are translated to the same Chinese word, "优雅" [2]. Though the two Chinese translations are generally correct, two distinct English words are conflated into only one Chinese word. This phenomenon is very common in translation. Many English sentiment words have identical or similar meaning. Corresponding to this meaning, there are also many possible translations in the target language, among which one translation is often dominant. As a result, when those English sentiment words are translated individually, this dominant translation are very likely to be picked out, whether by using bilingual dictionary or machine translation engine. In this circumstance, many translation variations are lost.

In order to recover the lost translation variations, we put the English words into different contexts. By using different contexts, we effectively prompt the machine translation engine to query the large scale parallel corpora that it is trained on, and then to return the most accurate translations in the target language. Furthermore, we can take advantage of the polysemy of words; one word can mean different things and it usually has various target language translations. Our context-based method effectively lead to translation diversity.

The flow chart of our approach is provided in Figure 1. As seen, we divide the overall process into three steps: (1) Generating the context; (2) Translation; (3) Extraction.

[2] All the following translation examples are obtained by using Google Translate (http://translate.google.com/)

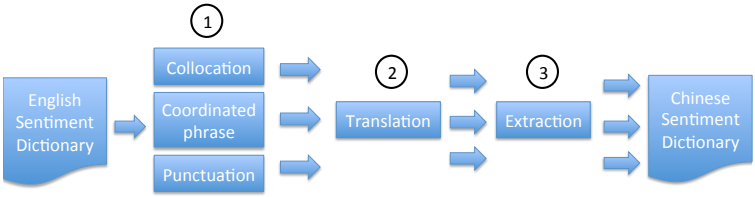

Figure 1: The Flow Chart of Our Approach

First, We devise the following three methods to generate contexts for translation.

- **Collocation**: We obtain the most frequent bi-grams containing the English word. This technique effectively makes the word meaning more specific and concrete, which helps the translation engine to pick out more accurate and diverse translations. For example, we generate "graceful voice" and "graceful dance". Given the contexts, "voice" and "dance", two "graceful" are translated to "优美" and "曼妙", respectively, which are more natural Chinese translations.
- **Coordinated phrase**: We combine two English words that have the same Chinese translations. This makes the translation engine less likely to return the same translations for both words. For example, we create a coordinated phrase by joining "elegant" and "graceful" with the word "and". Joining together, the translations for both words are different from the original translation. More interestingly, putting the two English words in different orders lead to different translations.
- **Punctuation**: We place a punctuation mark, such as period or question mark, at the end of the English word. We use this simple rule to limit the possible parts-of-speech of the translations. For example, "effusive." is translated to "热情洋溢", while "effusive" is translated to "感情奔放的"; after adding punctuation context, "effusive" is translated to words that have different parts-of-speech. We can also combine this technique with the coordinated phrase technique.

Concretely, We use a bi-gram language model for generating possible collocations. Instead of creating our own language model from large corpora, we leverage the Microsoft Web N-gram Services (Wang et al., 2010)[3], an online N-gram corpus that built from Web documents. We choose the bi-gram language model trained on document titles. Given each English polarity word w_1, we use the language model to generate up to the 1000 most frequent bi-grams w_1w_2.

To create coordinated phrases, we first translate all sentiment words using Google Translate. And then we create coordinated phrases for the English sentiment words which are translated into the same Chinese word. We select those English words and join them with the word "and". The punctuation context are generated by appending a period after the given English word. By combining and using both rules simultaneously, we can generate even more queries.

[3]http://research.microsoft.com/web-ngram

Lexicon	#POS	#NEG	#TOTAL
MPQA(EN)	1,481	3,080	4,561
DICT	742	1,139	1,881
DICT + Stem	814	1,230	2,044
DICT + Multiple	2,811	3,799	6,610
MT	1078	2,104	3,182
CONTEXT	**3,511**	**5,210**	**8,721**

Table 2: Vocabulary Size of Different Lexicons

Then we translate the resulted queries. We use Google Translate[4] as the online machine translation service. After that, we extract the foreign sentiment words from machine translation results. This step is language dependent but is often straight-forward. In this paper, we conduct experiment on Chinese. We first use Stanford Chinese Word Segmenter[5] for segmentation, and then use the position of the words and the punctuation between the words to locate the sentiment polarity word candidates. Finally we prune the candidates list by removing the words have less than 3 occurrences.

Discussion Our approach can be applied to construct sentiment dictionaries in other languages as well. Depending on the target language, we might need to make some small modifications. Word segmentation is unnecessary for most European languages. And in some languages, we need to consider the word order issues when extracting the sentiment words from the translation results, since translation engine might reorder the queries. For example, in Arabic, the modifying adjectives are placed before the nouns, which is different from English; and also in Arabic, the words are written from right to left.

3 Experimental Study

We use the MPQA subjective lexicon (Wilson et al., 2005) as the English lexicon. We only keep the strong subjective entries, which include 1,481 positive and 3,080 negative entries. For the purpose of comparison, we implemented the following baseline approaches. The first three baselines rely on a bilingual dictionary. We use the LDC (Linguistic Data Consortium) English-Chinese bilingual wordlists[6], which is also used in (Wan, 2008). This dictionary contains 18,195 entries. Each English entry is mapped to a list of Chinese words or expressions.

As shown in Table 2, the first baseline (DICT) looks up the English entry in the bilingual dictionary and use the first translation in the corresponding Chinese translation lists. Only 1,148 positive and 2,004 negative entries can be found in the bilingual dictionary, while about 1,500 entries are lost in the bilingual dictionary. After removing duplicate Chinese entries in the translated Chinese sentiment lexicon, only 742 unique positive entries and 1,139 negative entries remain. To improve the chances of finding English sentiment words in the bilingual dictionary, we use the Porter stemmer[7] to first obtain the lemmatization forms of the English sentiment words and then search them again in the bilingual dictionary. The results (DICT + Stem) show that the recall slightly improves, but the size

[4] http://translate.google.com/
[5] http://nlp.stanford.edu/software/segmenter.shtml
[6] http://projects.ldc.upenn.edu/Chinese/LDC_ch.htm
[7] http://tartarus.org/~martin/PorterStemmer/

Lexicon	Precision
DICT	93%
DICT + Stem	93%
DICT + Multiple	82%
MT	91%
CONTEXT	91.5%

Table 3: Precision of Different Lexicons

of the Chinese sentiment lexicon is still much smaller than the English sentiment lexicon. We further expand the sentiment lexicon by including all translations of each English entry, with the exception of the translations that contain punctuations and are longer than 6 characters; we filter translations longer than 6 characters since most of these sentiment words or phrases are merely the combinations of the shorter words and phrases. From the results of this baseline (DICT + Multiple), we can see that this approach can remarkably expand the lexicon. However, This method introduces many noises, as described later. Instead of using bilingual dictionary, we can use the machine translation engine to directly translate the English sentiment words. The results of this baseline (denoted by MT) show that the it is superior to the DICT baseline, but the vocabulary it covers is still too limited. The results of our approach (denoted by CONTEXT) are shown at the bottom. The lexicon generated are significantly larger than all other lexicons.

3.1 Lexicon Quality

To evaluate the precision of the sentiment lexicons generated by using our approach and the baselines, we sample 200 entries for each polarity (positive and negative) from each lexicon and compute their precision. Table 3 depicts the comparison, from which we can see that the positive lexicon generated by DICT + Multiple is very noisy. By contrast, our approach can generate a large lexicon with high precision. Though other lexicons have very high precisions, the vocabularies are too small.

To investigate why Dictionary-based translation methods lead to relatively low coverage lexicon, we look into the generated Chinese sentiment lexicon and identify three causes. First, the bilingual dictionary is not a comprehensive list of the Chinese translations of each English word. Instead it just includes a few translations to help people to understand the meaning of the English word. Second, the bilingual dictionary often translates different English words to one Chinese word. Third, the bilingual dictionary does not include translations of multi-word expressions. The MT baseline alleviates the problem of multi-word expressions, but it still suffers from the first two problems. We also study the noise words introduced by DICT + Multiple. Most of noise words are direct translations of one particular sense of some polarity English words. For example, “吸入”, which means “breathe in”, is included because the polarity word “inspire” has this sense as a technical term.

One other possible approach to enlarging the lexicon is to use N-best translations for English polarity words. We do not explore this approach in this paper for two reasons. First, online machine translation services often do not provide convenient interfaces for retrieving N-best translation results. Second, based on our observation, N-best translations of individual sentiment words are similar to multiple translations using a bilingual

Lexicon	NTCIR	Weibo
DICT	61.9%	57.6%
DICT + Stem	61.9%	57.5%
DICT + Multiple	64.7%	61.7%
MT	66.2%	64.6%
CONTEXT	**70.1%**	**73.5%**

Table 4: Classifier Accuracy Using Different Lexicons

dictionary. Both approaches tend to produce general and abstract words like “高兴” and “快乐” (both mean “happy”), but have difficulties in generating Chinese idioms such as “兴高采烈” , which also expresses “happy”, but in a more vivid way. One interesting fact of the CONTEXT lexicon is that it includes many four-characters idioms, which are widely used in Chinese but rarely found in bilingual dictionaries. By contrast, dictionary-based approaches often fail to generate those idioms.

3.2 Lexicon Usefulness in Sentiment Classification

One important application of sentiment lexicons is document sentiment classification, predicting whether a given document to express a positive or negative attitude. Sentiment lexicons can be used either as the basic resources for dictionary-based classifiers, or as a preprocessing step to generate augmented features for corpus-based classifiers. Therefore, we evaluate the usefulness of the lexicons by evaluating the performance of classifiers using different lexicons.

We use a dictionary-based sentiment classification approach. Besides the sentiment lexicon, we also use a negation lexicon, which collect the terms that can reverse the sentiment. The negation lexicon we use is the Chinese translation of negation lexicon from OpinionFinder[8]. The polarity score of each document is the sum of all the polarity of sentiment words in the document; if a negation word is in the context window of the sentiment word[9], we inverse the polarity of this sentiment word. If the overall polarity score is less than 0, we label this document as negative; otherwise the document is predicted as positive.

We test the classifiers on two data sets, which belong to different genres. The first test data set comes from the NTCIR Opinion Analysis Pilot Task data set (Seki et al., 2007, 2008). This data set contains 4,294 Chinese sentences, 2,378 being positive sentences and 1,916 being negative. Those sentences are all extracted from news. The second data set is collected from Weibo[10], a micro-blogging service website in China. We sample 5,000 messages from Weibo, and label them manually. To be consistent with the NTCIR data set, we only keep the positive and negative message. The resulting Weibo data set contains 906 positive messages and 807 negative messages. Each sentence/message is segmented into Chinese words by using Stanford Chinese word segmenter.

We report the results in Table 4. As seen, the classifier using our CONTEXT lexicon obtains the highest accuracy on both data sets. Comparing the results in the NTCIR and

[8]http://www.cs.pitt.edu/mpqa/opinionfinder.html
[9]We use a distance window of two words
[10]http://weibo.com

the Weibo column, it is interesting to note that the Weibo data set decreases the accuracy of classifiers using all lexicon but CONTEXT lexicon. As described in the previous section, our CONTEXT lexicon contains many Chinese idioms, which are seldom used in news. Hence our lexicon performs even better in user generated contents, such as blogs and user reviews.

We also note that bilingual dictionary is not an effective method for adapting resources cross-lingually, since classifiers with MT lexicon performs better than all the ones with DICT variants. Another interesting fact is that using larger lexicon do not always lead to better classifier accuracy; the classifier with MT lexicon performs better than the one with DICT + Multiple, despite the fact that the DICT + Multiple lexicon is much larger than the MT lexicon.

Conclusion and Future Work

In this paper, we propose an approach to leveraging publicly available machine translation services for creating sentiment lexicons from English sentiment lexicons. By placing English sentiment words in carefully crafted contexts, we effective prompt the translation engine to translate the same sentiment words differently. The experiment results show that our approach can obtain a high sentiment word coverage while achieving relatively high precision. This approach treats the machine translation engine as a black box. In the future, we will experiment with the ideas of directly using the underlying parallel corpus for creating sentiment lexicons.

Acknowledgment

This work was partially supported by National High Technology Research and Development Program of China (863 Program) (No. 2012AA011101), National Natural Science Foundation of China (No.91024009, No.60973053), the Specialized Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20090001110047), and 2009 Chiang Ching-kuo Foundation for International Scholarly Exchange (under Grant No. RG013-D-09).

References

Esuli, A. and Sebastiani, F. (2006). SentiWordNet: a publicly available lexical resource for opinion mining. In *Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation*, Genova, IT.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. The MIT press.

Hassan, A., Abu-Jbara, A., Jha, R., and Radev, D. (2011). Identifying the semantic orientation of foreign words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 592-–-597.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of KDD '04, the ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–--177, Seattle, US. ACM Press.

Kerry, L. V. and McDonald, H. R. (2010). The viability of web-derived polarity lexicons. In *Annual Conference of the North American Chapter of the ACL*.

Ku, L. and Chen, H. (2007). Mining opinions from the web: Beyond relevance retrieval. *Journal of the American Society for Information Science and Technology*, 58(12):1838–-1850.

Mihalcea, R., Banea, C., and Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 976.

Rao, D. and Ravichandran, D. (2009). Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 675–-682, Athens, Greece. Association for Computational Linguistics. ACM ID: 1609142.

Seki, Y., Evans, D. K., Ku, L.-W., Chen, H.-H., Kando, N., and Lin, C.-Y. (2007). Overview of opinion analysis pilot task at NTCIR-6. In *Proceedings of NTCIR-6 Workshop Meeting*, page 265–278.

Seki, Y., Evans, D. K., Ku, L.-W., Sun, L., Chen, H.-H., Kando, N., and Lin, C.-Y. (2008). Overview of multilingual opinion analysis task at NTCIR-7. In *Proc. of the Seventh NTCIR Workshop*.

Steinberger, J., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Lenkova, P., Steinberger, R., Tanev, H., Vuez, S., and Zavarella, V. (2012). Creating sentiment dictionariesvia triangulation. *Decision Support Systems*.

Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Wan, X. (2008). Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 553–-561, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wan, X. (2011). Bilingual co-training for sentiment classification of chinese product reviews. *Computational Linguistics*, 37(3):587–616.

Wang, K., Thrasher, C., Viegas, E., Li, X., and Hsu, B. (2010). An overview of microsoft web n-gram corpus and applications. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, page 45–48.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, CA.

Xu, G., Meng, X., and Wang, H. (2010). Build chinese emotion lexicons using a graph-based algorithm and multiple resources. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1209–-1217, Beijing, China. Coling 2010 Organizing Committee.