# An Approach to Cross-lingual Sentiment Lexicon Construction

Chia-Hsuan Chang, Ming-Lun Wu and San-Yih Hwang
Department of Information Management
National Sun Yat-sen University
Kaohsiung, Taiwan
sham82503@gmail.com

**Abstract—Lexicon-based sentiment analysis is a popular and practical approach for sentiment analysis. However, sentiment lexicons, which may be abundant in some language such as English, are scarce in many other languages. The cross-lingual lexicon learning aims to extend lexicons for the language with less resources from those lexicons available in other languages. In this paper, we propose an approach that builds a skip-gram variant to map word spaces across languages so as to construct lexicons for the language with less resources. We show in our preliminary experiment that our approach can generate lexicons that are similar to those crafted by human experts.**

***Keywords- text mining; sentiment analysis; cross-lingual lexicon learning; skip-gram; lexical relation***

## I. INTRODUCTION

Sentiment analysis aims to detect subjective and emotional information from texts and has been applied to a wide range of domains. The success of sentiment analysis relies on the existence of various resources, including domain specific lexicons, corpus, and ontologies. Cross-lingual sentiment analysis intends to utilize the rich resources available in some language, such as English, to facilitate the sentiment analysis of another language that has less resources, and several methods have been proposed [1]–[5]. Techniques for cross-lingual sentiment analysis can be classified into two main categories: corpus-based and lexicon-based methods [1]. Corpus-based methods leverage the sentiment corpus available in the resource-rich language. Most works of this genre employ neural networks to extract word features shared across multiple languages, and a classifier is subsequently constructed on top of those features [2]. However, corpus-based methods often suffer from inefficiency and lack of interpretability. Lexicon-based methods, on the other hand, rely on the cross-lingual sentiment lexicons for determining the sentiment polarity of a document, sentence, or entity. Such methods do not need sentiment corpus and has better interpretability. In this work, our focus is on lexicon-based methods.

There exist some sentiment lexicons carefully crafted by psychologists and linguists, e.g., LIWC [6], GI [7] and EmoLex [8]. Nevertheless, they are of limited number and only available in a few languages. To expand the lexicons that cover multiple languages, the idea of cross-lingual lexicon learning has been proposed. The cross-lingual lexicon learning can be more specifically defined as follows. Consider *n* languages $\{l_1, l_2, \ldots, l_n\}$, and each language has a corpus (without sentiment labeling) and optionally a lexical database in which some word relations such as synonyms and antonyms are given. Of the *n* languages, $l_1$, called dominating language, has a pre-defined sentiment lexicon $V_{l_1}^{category}$, where *category* could be positive, negative, fear, joy, etc. Take $V_{English}^{positive}$ as an example; it may contain positive words such as *happy*, *good*, and *excellent*. The goal of cross-lingual lexicon learning is to learn the corresponding lexicons for $V_{l_i}^{category}$, *i*=2…*n*, by leveraging corpus and lexical database of each language.

Several methods of cross-lingual lexicon learning have been proposed to infer the lexicons for languages of less resources by leveraging the rich resources available in the dominating language, namely English. It has been shown by Mihalcea et al. [3] that direct machine translation techniques generate lexicons of low coverage. Other researches use multilingual WordNet, e.g., [4], or parallel corpus, e.g., [5], to extend lexicons for languages with less resources. However, parallel corpus is difficult to come by in practice and WordNet-based approach often yields lexicons with low coverage. Our approach also utilizes WordNet for each language, specifically the synonym and antonym relations, with the help of corpus in each language to enhance coverage. In addition, we do not need parallel corpus to align words of different languages. The experiments show that our proposed approach can generated lexicons with good quality.

The remainder of this paper is structured as follows. In Section II, we present the details of leveraging lexical relations in our proposed approach and the mechanisms of learning lexicons across languages. Section III reports our preliminary experimental results.

## II. THE APPROACH

### A. Building Word Graph

We create two separate graphs for each language $l \in \{l_1, l_2, \ldots, l_n\}$, namely synonym graph, denoted $G_l^{\mathrm{Syn}} = (V_l, E_l^{\mathrm{Syn}})$ and antonym graph, denoted $G_l^{\mathrm{Ant}} = (V_l, E_l^{\mathrm{Ant}})$. In both graphs, $V_l$ is the set of vocabularies in language $l$, and $(v_{l,i}, v_{l,j}) \in E_l^{\mathrm{Syn}}$ ($E_l^{\mathrm{Ant}}$), $v_{l,i}$ and $v_{l,j} \in V_l$, if $v_{l,i}$ and $v_{l,j}$ are synonyms (antonyms) in language *l*.

### B. Incorporating Pseudo Contexts for Skip-gram

Skip-gram [9] is an efficient and well-studied method for training word representations from a large volume of text

978-1-7281-2772-9/19/$31.00 ©2019 IEEE
DOI 10.1109/BigDataCongress.2019.00030

dataset. The idea is to build word vector space model that predicts word's contexts at the sentence level. For training a skip-gram model, we need to prepare a set of training samples *Q*. Each sample $q \in Q$ is a pair (*w, k*), where each center word $w$ would pair with each context word $k$ which is within the given word distance from $w$ in some sentences contained in the corpus. The strength of skip-gram is that it is capable of generating word vectors that well preserve their semantics. That is to say, words with similar context tend to be close in their vectors. However, the resultant word vectors may not well distinguish the sentiment polarity between words, because words with similar context words do not necessarily possess the same sentiment polarity. For example, "joy" and "angry", despite their opposite sentiment, could be highly similar in their word vectors because they may share many common context words.

For mitigating this issue, we leverage the synonym and antonym relations from the existing lexical database, which serve as the pseudo contexts for skip-gram. Specifically, besides the existing contextual sample (*w, k*), we add the synonym samples $(w, s_1), (w, s_2), \dots, (w, s_n)$ and antonym samples $(w, a_1), (w, a_2), \dots, (w, a_m)$, where $s_i$ ( $a_j$ ) is a synonym (indirect antonym) of $w$. That is to say, a pseudo context word of a word *w* has similar meaning to *w* as determined by the lexical relations. To prevent pseudo contextual words from dominating the effect, we randomly choose synonym and antonym samples according to the following probability functions:

$$f_{syn}(w, s_i) = \frac{\rho}{\gamma^{\alpha}} \quad (1)$$

$$f_{ant}(w, a_j) = \begin{cases} \frac{\rho}{\delta^{\alpha}} & if\ \delta\ is\ even \\ 0 & otherwise, \end{cases} \quad (2)$$

where $\rho$ is a fraction number, $\gamma$ ($\delta$) is the length of shortest path between $w$ and $s_i$ $(a_j)$, and $\alpha \in \mathbb{Z}^+$ is an adjusting factor. For example, let $\rho$ and $\alpha$ be both 1. A direct synonym $s_i$ of *w*, i.e., $\gamma = 1$, will always be selected because $f_{syn}(w, s_i) = 1$. An indirect synonym $s_j$ of *w*, i.e., $\gamma = 2$, will have half the chance of being selected because $f_{syn}(w, s_j) = 0.5$. The number of pseudo contextual samples can be adjusted by changing the value of $\rho$. The lower of $\rho$, the less number of pseudo contextual samples. Notice that $E_l^{\text{Ant}}$ is constructed by antonym relation; therefore, only the node pair whose length of shortest path is even will be considered because they may possess similar meaning. By adding the pseudo contexts, the skip-gram starts to consider the lexical relations.

The objective function of the skip-gram variant is shown in Equation (3). Notice that by adding pseudo contexts, each training sample $q$ now is denoted (*w, c*), where *c* could be contextual word *k* or pseudo-contextual word $s_i$ $(a_j)$.

$$\underset{\theta}{\text{argmax}} \sum_{q \in Q} log p(c|w), \quad (3)$$

where $p(c|w)$ is the probability of $w$ for predicting its context $c$. For improving the training efficiency, we approximate the $p(c|w)$ by negative sampling [10]:

$$p(c|w) = log\sigma(u^T(c)v(w)) + m\mathbb{E}_{\tilde{c} \sim P_{noise}} log\sigma(-u^T(\tilde{c})v(w)), \quad (4)$$

where $u^T(c)$ represents a $d$-dimensional output vector of the context word $c$, $v(w)$ is the *d*-dimensional word vector of the target word $w$, $m$ is the number of negative samples and $\tilde{c}$ is the negative sample drawn from modified unigram distribution $P_{noise}$ [10]. As a result, we apply stochastic gradient descent (SGD) method to obtain the optimal $u$ and $v$.

### *C. Learning Lexcion by Translation Matrix*

We adopt the linear transformation method [11] to learn a translation matrix that maps the word vector space of a non-dominating language to that of the dominating language. The reason of choosing linear transformation method is its simplicity and competitive performance with other alternative methods [12]. In the linear transformation, we only need a small set of bilingual dictionary entries $\{v_s(i), v_t(i)\}_{i=1}^{N}$ to train the translation matrix $\Omega \in \mathbb{R}^{d \times d}$, where $v_s(i) \in \mathbb{R}^d$ denotes the word vector of $i$-th word in source (non-dominating) language and $v_t(i) \in \mathbb{R}^d$ is the word vector of its translation in target (dominating) language. The training objective of Ω is shown in Equation (5).

$$\underset{\Omega}{argmin} \sum_{i=1}^{N} \left|\left| \Omega \cdot v_s(i) - v_t(i) \right|\right|^2. \quad (5)$$

After obtaining Ω, we can compare words across languages in the word space of the dominating language. As a result, we can start to learn sentiment lexicons $V_{\{l_2,\dots,l_n\}}^{\text{category}}$ by taking $V_{l_1}^{\text{category}}$ as queries. More precisely, words in language other than $l_1$ whose cosine similarity to some word in $V_{l_1}^{\text{category}}$ surpasses a threshold $\tau$, will be added to $V_{\{l_2,\dots,l_n\}}^{\text{category}}$.

## III. PRELIMINARY EVALUATION

We use Reuters Corpus Volume 2 (RCV2) [13] in our experiments, which contains news articles in thirteen languages. Only Chinese and English versions are used in our experiments, each containing 24K articles. For synonym and antonym relations, we consult the Open Multilingual Wordnet (OMW) [14]. For the sentiment lexicon of the dominating language, namely English, we use NRC emotion lexicon, EmoLex [8]. Finally, we use the Chinese-English dictionary in Facebook MUSE project[1] as the seeds for aligning the word space of Chinese and English. In those dictionary entries, 5,381 pairs appear in our corpus.

We use the same settings of skip-gram for word space construction in both languages, where the size of sliding window is 5, number of negative samples is 64, dimension of word vectors is 128, and number of epochs is 10. To evaluate

[1] https://github.com/facebookresearch/MUSE#ground-truth-bilingual-dictionaries

the performance of the learned Chinese lexicons, we compare them to Chinese LIWC (C-LIWC) [15]. In particular, we apply C-LIWC and the learned lexicons separately on each article in RCV2 Chinese corpus and record the word count of each category. Pearson's correlation between a learned lexicon and the C-LIWC is computed. Higher correlation indicates that the learned lexicon is closer to the benchmark dictionary and therefore better.

Due to the space limitation, we only report the performance result when $\alpha = \gamma = 3$ because they yield the best results. It indicates that similar words that are two or three steps away in synonym and antonym graphs are unlikely to be included. Fig. 1 shows how the value of $\rho$ affects the performance. It can be seen that in our experiment it is better to set higher value for $\rho$, which indicates more pseudo context words contribute to better performance. We then show the impact of incorporating synonyms and antonyms into our approach in Fig. 2. By proposing pseudo contexts, our approach outperforms the original skip-gram in learning sentiment lexicons by 11% in mean correlation. We also display some sample results in Table 1. It shows our approach can filter out the irrelevant sentiment words and makes it more appropriate for learning sentiment lexicon when compared to pure skip-gram method.

Table 1. The top three similar Chinese words of "Decrease" and "Rise" by applying skip-gram and our approach

| Word in source language | Method | Top 3 similar words in target language |
|---|---|---|
| Decrease | Skip-gram | 降幅、德西年、同期 |
| | Our Approach | 減少、暴跌、降低 |
| Rise | Skip-gram | 泰唔士報、已盡、下降 |
| | Our Approach | 升起、上漲、上升 |

## IV. Conclusion and Future Work

We have proposed an approach that learns sentiment lexicons for a language with less resources by leveraging the rich resources pertaining to English. The proposed approach has demonstrated promising result in our preliminary experiment. In the future, we will further enhance our method to increase the vocabulary coverage and the accuracy of lexicons.

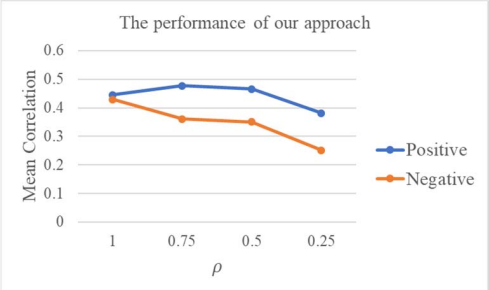


Figure. 1. The mean correlation of positive and negative lexicons with Chinese LIWC with different value of $\rho$

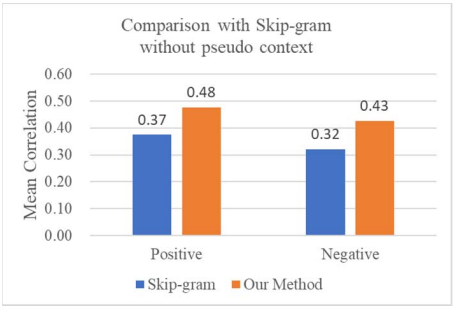


Figure. 2. The mean correlation of positive and negative lexicons with Chinese LIWC when applying our approach and skip-gram


## Acknowledgment

This paper is supported in part by the ministry of science and technology in Taiwan under grant number MOST 106-2410-H-110-017-MY3.


## References


[1] K. Dashtipour *et al.*, "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques," *Cogn. Comput*, vol. 8, pp. 757–771, 2016.

[2] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger, "Adversarial deep averaging networks for cross-lingual sentiment classification," *Trans. Assoc. Comput. Linguist.*, vol. 6, pp. 557–570, 2018.

[3] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *Proceedings of the 45th annual meeting of the association of computational linguistics*, 2007, pp. 976–983.

[4] A. Hassan and A. Abu-Jbara, "Identifying the Semantic Orientation of Foreign Words," in *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, 2011, pp. 592–597.

[5] D. Gao, F. Wei, W. Li, X. Liu, and M. Zhou, "Cross-lingual Sentiment Lexicon Learning With Bilingual Word Graph Label Propagation," *Comput Linguist*, vol. 41, no. 1, pp. 21–40, 2015.

[6] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Soc. Psychol.*, vol. 29, no. 1, pp. 24–54, 2010.

[7] P. J. Stone, D. C. Dunphy, and M. S. Smith, "The general inquirer: A computer approach to content analysis.," 1966.

[8] S. M. Mohammad and P. D. Turney, "Crowdsourcing a Word-Emotion Association Lexicon," *Comput Intell*, vol. 29, no. 3, pp. 436–465, 2013.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *ArXiv Prepr. ArXiv13013781*, 2013.

[11] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *ArXiv Prepr. ArXiv13094168*, 2013.

[12] S. L. Smith, D. H. Turban, S. Hamblin, and N. Y. Hammerla, "Offline bilingual word vectors, orthogonal transformations and the inverted softmax," *ArXiv Prepr. ArXiv170203859*, 2017.

[13] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," *J. Mach. Learn. Res.*, vol. 5, no. Apr, pp. 361–397, 2004.

[14] F. Bond and R. Foster, "Linking and extending an open multilingual wordnet," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, vol. 1, pp. 1352–1362.

[15] C.-L. Huang *et al.*, "The development of the Chinese linguistic inquiry and word count dictionary.," *Chin. J. Psychol.*, 2012.