IPASJ International Journal of Information Technology (IIJIT)
Web Site: http://www.ipasj.org/IIJIT/IIJIT.htm
A Publisher for Research Motivation ..... Email:editoriijit@ipasj.org
Volume 5, Issue 11, November 2017 ISSN 2321-5976

Cross-lingual Sentiment Lexicon Learning With Bilingual Word Graph Label Propagation

Hamilton, Alexander

Erasmus University Rotterdam

ABSTRACT

In this article we tend to address the task of cross-lingual sentiment lexicon learning, that aims to mechanically generate sentiment lexicons for the target languages with obtainable English sentiment lexicons. we tend to formalize the task as a learning drawback on a bilingual word graph, within which the intra-language relations among the words within the same language and therefore the lingua franca relations among the words between totally different languages area unit properly drawn. With the words within the English sentiment lexicon as seeds, we tend to propose a bilingual word graph label propagation approach to induce sentiment polarities of the untagged words within the target language. notably, we tend to show that each equivalent word and opposite word relations is wont to build the intra-language relation, which the word alignment info derived from bilingual parallel sentences is effectively leveraged to make the inter-language relation. The analysis of Chinese sentiment lexicon learning shows that the projected approach outperforms existing approaches in each exactness and recall. Experiments conducted on the NTCIR information set any demonstrate the effectiveness of the learned sentiment lexicon in sentence-level sentiment classification.

1. INTRODUCTION

A sentiment lexicon is thought to be the foremost valuable resource for sentiment analysis (Pang and Lee 2008), and lays the groundwork of abundant sentiment analysis analysis, as an example, sentiment classification (Yu and Hatzivassiloglou 2003; Kim and Hovy 2004) and opinion summarisation (Hu and Liu 2004). To avoid manually expanding upon sentiment words, Associate in Nursing mechanically learning sentiment lexicon has attracted extended attention within the community of sentiment analysis. the present work determines word sentiment polarities either by the applied mathematics info (e.g., the co-occurrence of words with predefined sentiment seed words) derived from an outsized corpus (Riloff, Wiebe, and Wilson 2003; Hu and Liu 2006) or by the word linguistics info (e.g., equivalent word relations) found in existing human-created resources (e.g., WordNet) (Takamura, Inui, and Okumura 2005; Rao and Ravichandran 2009). However, current work primarily focuses on English sentiment lexicon generation or enlargement, whereas sentiment lexicon learning for alternative languages has not been well studied. during this article, we tend to address the problem of cross-lingual sentiment lexicon learning, that aims to get sentiment lexicons for a non-English language (hereafter observed as "the target language") with the assistance of the obtainable English sentiment lexicons. The underlying motivation of this task is to leverage the present English sentiment lexicons and substantial linguistic resources to label the sentiment polarities of the words within the target language. to the current finish, we want Associate in Nursing approach to transferring the sentiment info from English words to the words within the target language. The few existing approaches initial build word relations between English and therefore the target language. Then, supported the word relation and English sentiment seed words, they confirm the sentiment polarities of the words within the target language. In these 2 steps, relation-building plays a basic role as a result of it's answerable for the transfer of sentiment info between the 2 languages. 2 approaches area unit typically wont to connect the words in several languages within the literature. One relies on translation entries in cross-lingual dictionaries (Hassan et al. 2011). the opposite depends on a MT (MT) engine as a recording machine to translate the sentiment words in English to the target language (Steinberger et al. 2011). the 2 approaches in Duh, Fujino, and Nagata (2011) and Mihalcea, Banea, and Weibe (2007) tend to use atiny low set of vocabularies to translate the language, that results in an occasional coverage of generated sentiment lexicons for the target language. to unravel this drawback, we tend to propose a generic approach to addressing the task of cross-lingual sentiment lexicon learning. Specifically, we tend to model this task with a bilingual word graph, that consists of 2 intra-language subgraphs Associate in Nursingd an lingua franca subgraph. The intra-language subgraphs area unit wont to model the linguistics relations among the words within the same languages. once building them, we tend to incorporate each equivalent word and opposite word relations in an exceedingly novel manner, drawn by positive and negative sign weights within the subgraphs, severally. These 2 intra-language subgraphs area unit then connected by the inter-language subgraph. we tend to propose Bilingual word graph Label Propagation (BLP), that at the same time takes the inter-language relations and therefore

the intra-language relations under consideration in Associate in Nursing repetitive manner. Moreover, we tend to leverage the word alignment info derived from a parallel corpus to make the inter-language relations. we tend to connect 2 words from totally different languages that area unit aligned to every alternative in an exceedingly parallel sentence try. Taking advantage of an outsized parallel corpus, this approach considerably improves the coverage of the generated sentiment lexicon.

We build the subsequent contributions during this article.

1. we tend to gift a generic approach to mechanically learning sentiment lexicons for the target language with the obtainable sentiment lexicon in English, and that we formalize the cross-lingual sentiment learning task on a bilingual word graph.
2. we tend to build a bilingual word graph by victimisation equivalent word and opposite word relations and propose a bilingual word graph label propagation approach, that effectively leverages the inter-language relations and each varieties (synonym and antonym) of the intra-language relations in sentiment lexicon learning.
3. we tend to leverage the word alignment info derived from an outsized range of parallel sentences in sentiment lexicon learning. we tend to build the inter-language relation within the bilingual word graph upon word alignment, and attain vital results.

2. CONNECTED WORK

2.1 English Sentiment

Lexicon Learning normally, the work on sentiment lexicon learning focuses primarily on English and might be classified as co-occurrence–based approaches (Hatzivassiloglou and McKeown 1997; Riloff, Wiebe, and Wilson 2003; Qiu et al. 2011) and semantic-based approaches (Mihalcea, Banea, and Wiebe 2007; Takamura, Inui, and Okumura 2005; Kim and Hovy 2004).

The co-occurrence-based approaches confirm the sentiment polarity of a given word in keeping with the applied mathematics info, just like the co-occurrence of the word to predefined sentiment seed words or the co-occurrence to product options. The applied mathematics info is principally derived from bound corpora. one in all the earliest work conducted by Hatzivassiloglou and McKeown (1997) assumes that the conjunction words will convey the polarity relation of the 2 words they connect. as an example, the conjunction word and tends to link 2 words with a similar polarity, whereas the conjunction word however is probably going to link 2 words with opposite polarities. Their approach solely considers adjectives, not nouns or verbs, and it's unable to extract adjectives that aren't joint by conjunctions. Riloff et al. (2003) outline many pattern templates and extract sentiment words by 2 bootstrapping approaches. Turney and Littman (2003) calculate the pointwise mutual info (PMI) of a given word with positive and negative sets of sentiment words. The sentiment polarity of the word is decided by average PMI values of the positive and negative sets. to get PMI, they supply queries (consisting of the given word and therefore the sentiment word) to the program. the quantity of hits and therefore the position (if the given word is close to the sentiment word) area unit wont to estimate the association of the given word to the sentiment word. Hu and Liu (2004) analysis sentiment word learning on client reviews and that they assume that the sentiment words tend to be related with product options. The frequent nouns and noun phrases area unit treated as product options.

The semantic-based approaches confirm the sentiment polarity of a given word in keeping with the word linguistic relation, just like the synonyms of sentiment seed words. The word linguistic relation is sometimes obtained from dictionaries, as an example, WordNet. one Kim and Hovy (2004) assume that the synonyms of a positive (negative) word area unit positive (negative) and its antonyms area unit negative (positive). Initializing with a collection of sentiment words, they expand sentiment lexicons supported these 2 types of word relations. Kamps et al. (2004) build a equivalent word graph in keeping with the equivalent word relation (synset) derived from WordNet. The sentiment polarity of a word is calculated by the shortest path to 2 sentiment words sensible and unhealthy. However, the shortest path cannot exactly describe the sentiment orientation, considering there area unit solely 5 steps between the word sensible and therefore the word unhealthy in WordNet (Hassan et al. 2011). Takamura et al. (2005) construct a word graph with the gloss of WordNet. Words area unit connected if a word seems within the gloss of another. The word sentiment polarity is decided by the load of its connections on the word graph. supported WordNet, Rao and Ravichandran (2009) exploit many graph-based semi-supervised learning ways like Mincuts and Label Propagation. The word polarity orientations area unit evoked by initializing some sentiment seed words within the WordNet graph. Esuli et al. (2006, 2007) and Baccianella et al. (2010) treat sentiment word learning as a machine learning drawback, that is, to classify the polarity orientations of the words in WordNet. They choose seven positive words and 7 negative words and expand them through the see-also and opposite relations in WordNet. These expanded words area unit then

used for coaching. They train a ternary classifier to predict the sentiment polarities of all the words in WordNet and use the glosses (textual definitions of the words in WordNet) because the options of classification.

2.2 Cross-Lingual Sentiment Lexicon Learning

The work on cross-lingual sentiment lexicon learning remains at Associate in Nursing early stage and might be classified into 2 varieties, in keeping with however they bridge the words in 2 languages. Mihalcea et al. (2007) generate sentiment lexicon for Romanian by directly translating nation sentiment words into Romanian through bilingual English–Romanian dictionaries. once grappling multiword translations, they translate the multiwords word by word. Then the valid translations should occur a minimum of 3 times on the online. The approach projected by Hassan et al. (2011) learns sentiment words supported English WordNet and WordNets within the target languages (e.g., Hindi and Arabic). Crosslingual dictionaries area unit wont to connect the words in 2 languages and therefore the polarity of a given word is decided by the typical touching time from the word to nation sentiment word set. These approaches connect words in 2 languages supported crosslingual dictionaries.

To improve the sentiment classification for the target language, Banea, Mihalcea, and Wiebe (2010) translate nation sentiment lexicon into the target language victimisation Google Translator.3 equally, Google Translator is employed by Steinberger et al. (2011). They manually turn out 2 high-level gold-standard sentiment lexicons for 2 languages (e.g., English and Spanish) then translate them into the third language (e.g., Italian) via Google Translator. They believe that those words within the third language that seem in each translation lists area unit possible to be sentiment words. These approaches connect the words in 2 languages supported MT engines. the most concern of those approaches is that the low overlapping between the vocabularies of natural documents and therefore the vocabularies of the documents translated by MT engines (Duh, Fujino, and Nagata 2011; Meng et al. 2012a). The defect of those MT-based approaches inevitably results in low coverage.

Our task resembles the task of cross-lingual sentiment classification, like Wan (2009), Lu et al. (2011), and Meng et al. (2012a), that classifies the sentiment polarities of product reviews. Generally, these studies use semi-supervised learning approaches and regard translations from tagged English sentiment reviews because the coaching information. The terms in every review area unit leveraged because the options for coaching, that has verified to be effective in sentiment classification (Pang and Lee 2008). we will regard the task of sentiment lexicon learning as word-level sentiment classification. However, for wordlevel sentiment classification, it's not simple to extract options for one word. while not decent options, it's troublesome for these approaches to perform well in learning. Another line of cross-lingual sentiment classification uses Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) or its variants, like Boyd-Graber and Resnik (2010) or He, Alani, and Chow dynasty (2010). These studies assume that every review may be a mixture of sentiments and every sentiment may be a chance over words. Then they apply the LDA-like approach to model the sentiment polarity of every review. however, this assumption might not be applicable in sentiment lexicon learning as a result of one word is thought to be the tokenish linguistics unit, and it's troublesome, if not not possible, to infer the latent topics from one word. Recall that totally different from the sentiment classification of product reviews wherever the instances area unit commonly freelance, words in sentiment lexicon learning area unit extremely connected with one another, like synonyms and antonyms. Through these relations, the words will naturally kind a word graph. so we tend to use the graph-based learning approach to leverage the word distributions in sentiment lexicon learning. within the next section, we'll introduce our projected graph-based cross-lingual sentiment lexicon learning.

3. CROSS-LINGUAL SENTIMENT LEXICON LEARNING

In this work, we tend to model the task of cross-lingual sentiment lexicon learning with a bilingual word graph, wherever (1) the words within the 2 languages area unit drawn by the nodes in 2 intra-language subgraphs, severally; (2) the equivalent word and opposite word relations at intervals every language area unit drawn by the positive and negative sign

3.1 Bilingual Word Graph Building With Parallel Corpus and Word Alignment

We represent the words in English and within the target language because the nodes of the bilingual word graph. we tend to use the equivalent word and opposite relations of the words within the same language to make W and Wein the intra-language graph, respectively. within the remainder of this section, we'll specialize in a way to build the inter-language relation. Intuitively, there area unit 2 ways in which to attach the words in 2 languages. One is to insert links to the words if there exist entry mappings between the words in bilingual dictionaries (e.g., the English–Chinese

dictionary). This technique is straightforward and simple, however it suffers from 2 limitations. (1) Dictionaries area unit static throughout an explicit amount, whereas the sentiment lexicon evolves over time. (2) The entries in dictionaries tend to be the expressions of formal and written languages, however individuals like victimisation conversational language in expressing their sentiments or opinions on-line. These limitations result in the low coverage of the links from English to the target language. another manner is to use Associate in Nursing MT engine as a recording machine to make the inter-language relation.

The word alignment info encodes the wealthy association info between the words from the 2 languages. we tend to area unit thus actuated to leverage the parallel corpus and word alignment to make the bilingual word graph for cross-lingual sentiment lexicon learning. we tend to take the words from each languages within the bilingual parallel corpus because the nodes within the bilingual word graph, and build the inter-language relations by connecting the 2 words that area unit aligned along in an exceedingly sentence try from a parallel corpus. There area unit many blessings of employing a parallel corpus to make the inter-language subgraph. First, giant parallel corpora area unit extensively used for coaching applied mathematics MT engines and might be simply reused in our task.

The parallel sentence pairs area unit sometimes mechanically collected and mined from the online. As a result, they contain the various and sensible variations of words and phrases embedded in sentiment expressions. Second, the parallel corpus is dynamically modified once necessary as a result of it's comparatively straightforward to gather from the online. Consequentially, the novel sentiment info inferred from the parallel corpus will simply update the present sentiment lexicons. These blessings will greatly improve the coverage of the generated sentiment lexicon, as incontestible later in our experiments.

3.2 Bilingual Word Graph Label Propagation

As usually used semi-supervised approaches, label propagation (Zhu and Ghahramani 2002) and its variants (Zhu, Ghahramani, and Lafferty 2003; Chow dynasty et al. 2004) are applied to several applications, like part-of-speech tagging (Das and Petrov 2011; Li, Graca, and Taskar 2012), image annotation (Wang, Huang, and peal 2011), supermolecule operate prediction (Jiang 2011; Jiang and McQuay 2012), then forth.

The underlying plan of label propagation is that the connected nodes within the graph tend to share a similar sentiment labels. In bilingual word graph label propagation, the words tend to share same sentiment labels if they're connected by equivalent word relations or word alignment and have a tendency to belong to totally different sentiment labels if connected by opposite relations.

4. EXPERIMENT

4.1 information Sets

We conduct experiments on Chinese sentiment lexicon learning. As in previous work (Baccianella, Esuli, and Sebastiani 2010), the sentiment words normally verbalizer lexicon area unit hand-picked because the English seeds (Stone 1997). From the GI lexicon we tend to collect a pair of,005 positive words and one,635 negative words. to make the bilingual word graph, we tend to adopt the Chinese–English parallel corpus, that is obtained from the news articles printed by Xinhua press agency in Chinese and English collections, victimisation the automated parallel sentence identification approach (Munteanu and Marcu 2005). Altogether, we tend to collect over 25M parallel sentence pairs in English and Chinese. we tend to take away all the stopwords in Chinese and English (e.g., (of) and am) in conjunction with the low-frequency words that occur fewer than five times. when preprocessing, we tend to finally have over 174,000 English words, among that three,519 words have sentiment labels and over 146,000 Chinese words that we want to predict the sentiment labels.

We initial generate each positive and negative scores for every untagged word then confirm the word sentiment polarities supported its scores. we tend to rank the 2 sets of new tagged sentiment words in keeping with their polarity scores. The top-ranked Chinese words area unit shown in Table one. we tend to manually label the top-ranked 1K sentiment words. For P@10K, 10dency to|we tend to} consecutive divide the highest 10K hierarchical list into ten equal components. 100 sentiment words area unit indiscriminately hand-picked from every half for labeling. the same as the analysis of TREC web log Distillation (Ounis, Macdonald, and Soboroff 2008), all the tagged words from every approach area unit utilized in the analysis. we tend to then valuate the hierarchical lists with 2 metrics, Precision@K and Recall.

In these approaches, μ is about to zero.1 as in Chow dynasty et al. (2004). The exactness of those approaches area unit shown in Figure three. The figure shows that the approaches supported the bilingual word graph considerably outmatch the one supported the monolingual word graph. The bilingual word graph will herald additional word relations and accelerate the sentiment propagation. Besides, within the bilingual word graph, nation sentiment seed words will frequently offer correct sentiment info. so we tend to observe the rise within the approaches supported the bilingual word graph in term of each exactness and recall (Table 2). Meanwhile, we discover that adding the opposite relation within the bilingual word graph slightly enhances exactness in top-ranked words and similar findings area unit discovered in our later experiments. It seems that the opposite relations depict word relations in an exceedingly additional correct manner and might refine the word sentiment scores additional exactly. However, the equivalent word relation and word alignment relation dominate, whereas the opposite relation accounts for less than atiny low proportion of the graph. it's onerous for the opposite relevance introduce new relations into the graph and so it cannot facilitate to any improve recall.

5. CONCLUSIONS AND FUTURE WORK

In this article, we tend to studied the task of cross-lingual sentiment lexicon learning. we tend to designed a bilingual word graph with the words in 2 languages and connected them with the inter-language and intra-language relations. we tend to projected a bilingual word graph label propagation approach to convert the sentiment info from English sentiment words to the words within the target language. The equivalent word and opposite relations among the words within the same languages area unit leveraged to make the intra-language relations. Word alignment derived from an outsized parallel corpus is employed to make the inter-language relations. Experiments on Chinese sentiment lexicon learning demonstrate the effectiveness of the projected approach. There area unit 3 main conclusions from this work. First, the bilingual word graph is appropriate for sentiment info transfer and therefore the projected approach will iteratively improve the exactness of the generated sentiment lexicon. Second, building the inter-language relations with the big parallel corpus will considerably improve the coverage. Third, by incorporating the opposite relations into the bilingual word graph, the BLP approach can do Associate in Nursing improvement in exactness. within the future, we'll explore the chance of increasing or generating the sentiment lexicons for multiple languages by bootstrapping.