Mining Market Trend from Blog Titles Based on Lexical Semantic Similarity

Fei Wang and Yunfang Wu

Institute of Computational Linguistic, Peking University, Beijing, China
sxjzwangfei@163.com, wuyf@pku.edu.cn

Abstract. Today blog has become an important medium for people to post their ideas and share new information. And the market trend of pricing Up/Down always draws people's attention. In this paper, we make a thorough study on mining market trend from blog titles in the field of housing market and stock market, based on lexical semantic similarity. We focus on the automatic extraction and construction of Chinese Up/Down verb lexicon, by using both Chinese and Chinese-English bilingual semantic similarity. The experimental results show that verb lexicon extraction based on semantic similarity is of great use in the task of mining public opinions on market trend, and that the performance of applying English similar words to Chinese verb lexicon extraction is well compared with using Chinese similar words.

Keywords: market trend, verb lexicon extraction, semantic similarity.

1 Introduction

The market trend of pricing Up/Down always draws people's attention. With the rapid development and expansion of the Internet, blog has been an important medium for people to post their ideas and share new information, and the titles of blogs make a concentrated presentation of the texts. In this paper, we try to mine market trend from blog titles, taking housing and stock market as two examples. Considering the guiding and recapitulative function of blog titles, we study the titles instead of the main body of the blogs to get the holders' opinion about market trend. The blog titles use a relatively restricted vocabulary, which makes the lexicon extraction proposed in this paper work well. But on the other hand, the expressions in titles can also be diversified and full of figurative meanings, adding the hardship of the task.

We would like to declare that the "market trend" expressed in this paper is different from traditional "opinion mining". In a traditional opinion mining system, opinions are classified as Positive/Negative/Neutral, which imply that the target invokes a positive/negative/neutral feeling. In this paper, the market trend is defined as Up and Down, and we try to mine people's idea (Up/Down) toward near-term direction of the market, instead of the emotional impact that the market trend may have on the public. For instance, the Up trend in housing market may invoke positive feeling for house

A. Gelbukh (Ed.): CICLing 2012, Part II, LNCS 7182, pp. 261–273, 2012.

holders while negative feeling for house buyers. In sum, the opinion mining in previous work is subjective, while the market trend in our paper is objective.

The following two examples explain our work:

(1) 北京/房价 /上涨/无 /绝期

Beijing/housing price/rise/without/end
Housing prices in Beijing will never stop increasing.

(2) 三 /大 /利空 /袭击 /大盘

Three/big/bad news/attack/broad market
Three negative news hit the broad market.

Example (1) tells that the house price trends Up, and (2) tells that the stock market trends Down.

A lexicon-based approach is applied to calculate the market trend of a given title, where verb lexicon plays a pivotal role (e.g., "上涨/rise" in Example (1) and "袭击/attack" in Example (2)). Realized the importance of verb lexicon, we focus on: 1) verb lexicon extraction based on distributional semantic similarity; 2) improving Chinese lexicon extraction by introducing English lexical semantic similarity. We conduct experiments using nine methods together, including two baselines, three individual semantic similarity methods and four ensemble methods. The experimental results validate our approach, and provide an easy but a promising way to use cross-lingual knowledge in distributional semantic similarity computation.

The remainder of this paper is organized as follows. In the next section, previous related work is discussed. Section 3 presents the lexicon-based approach and addresses the necessity of verb lexicon extraction. Section 4 describes our verb lexicon extraction methods in detail. Section 5 presents the experimental results and gives an analysis. Finally, Section 6 concludes this paper and proposes future work.

2 Related Work

2.1 Opinion Mining

Recently there has been extensive research in opinion mining, for which Pang and Lee (2008) give an in-depth survey of literature. Most work concerning opinion mining mainly runs on relatively complete text blocks, and a Positive/Negative tag is then attached to a given document. With the popularity of blogs, opinion mining on blog texts has attracted researchers' attention, such as the work of Chesley et al. (2006), Liu et al. (2010) and Park et al. (2010). There are only a few systems running on titles. Peramunetilleke and Raymond (2002) investigate how news headlines can be used to forecast intraday currency exchange rate movement. This paper focuses on blog titles, which is more difficult than main texts due to the diversified expressions and incomplete syntactic structures. What's more, the designed task of market trend in this paper is different from opinion mining in previous sentiment analysis work.

2.2 Similarity-Based Lexicon Extraction

Lexical semantic similarity has been widely applied to lexicon extraction. Previous work concentrates on entity extraction (Pennacchiotti and Pantel, 2009; Pantel et. al., 2009; Chaudhuri et. al., 2009), which is to extract instances of semantic classes (e.g., "*Ziyi Zhang*" and "*Li Gong*" are instances of the class *Actors*). Compared with noun lexicon extraction, less effort has been devoted to verb lexicon extraction. In the work of Shi et al. (2010), the performance of verb extraction is much worse than noun extraction while using semantic similarity. In this paper, we exploit semantic similarity to verb lexicon extraction in the task of mining market trend from blog titles, demonstrating the effectiveness of the application usage of verb extraction.

2.3 Cross-Lingual Knowledge in Semantic Similarity Computation

Most work concerning similarity-based lexicon extraction exploits only the knowledge of one language. In the field of synonym extraction, which is a small set of similar words, some studies have used cross-lingual knowledge. Wu and Zhou (2003) extract synonyms with a bilingual English-Chinese corpus, where the feature vector of a word is constructed by the translations and translation probabilities. Lin et al. (2003) identify synonyms among distributional similar words using bilingual dictionaries. Plas and Tiedemann (2006) find synonyms using automatic word alignment of parallel corpora, achieving much higher precision and recall scores than the monolingual syntax-based approach. In the field of Wikipedia-based semantic relatedness computation, which is different from distributional similarity computation in assumptions and methods, Sorg and Cimiano (2008), Potthast et al (2008) provide a new idea of using cross-lingual links/alignment of Wikipedia to map the explicit semantic analysis (ESA) vectors between different languages. In this paper, we extend the cross-lingual property from synonyms to semantic similar words, and provide a more fast way to make use of the full-fledged English resources in Chinese lexicon extraction.

3 Lexicon-Based Analysis

In this task, we adopt a simple but widely used lexicon-based approach. Our focus is not the algorithm itself but the impact of the lexicon, that is, we use this simple algorithm to examine the automatically extracted Up/Down verb lexicon.

3.1 The Lexicon-Based Algorithm

The algorithm is displayed in Table 1. Obviously, the performance of this approach relies heavily on the Up/Down lexicon, which is just the focus of this paper. The negation window is set to q=3, according to the empirical analysis.

Table 1. Algorithm of computing market trend

1. Each title T^i is segmented into a word set $T^i = w_1^i, w_2^i \ldots w_n^i$;
2. For each $w_j^i (1 \le j \le n)$, look it up in Up/Down lexicon and get $Trend(w_j^i)$;
3. Within the window of q words previous to w_j^i, if there is a negation term w′, $Trend(w_j^i) = -Trend(w_j^i)$;
4. Calculate the trend value expressed in title T^i: $Trend(T^i) = \sum_{1 \le j \le n} Trend(w_j^i)$;
5. Determine the trend tag $Tag(T^i)$: $Tag(T^i) = \begin{cases} Up \ \ if \ \ Trend(T^i) > 0 \\ Down \ \ if \ \ Trend(T^i) < 0 \\ unknown \ if \ \ Trend(T^i) = 0 \end{cases}$

3.2 The Verb Lexicon

In order to decide what kind of word is more important in blog titles, we make a statistics analysis on POS distribution on our data (the collection of our data will be described in Section 5.1). The results show that verb is the most frequently used type of words in titles, which accounts for 26.7% in housing market and 32.8% in stock market. In addition, verbs play a pivotal role in determining the meaning of sentences in Chinese language. So we regard verb as the main information-loader in titles. Thus, in the following section, we focus on the verb lexicon extraction.

We extracted all the verbs in the titles to constitute the candidate verb set, from which we filtered out the directional verbs and dummy verbs (e.g., 加以/do, 进行/ do), according to the Grammatical Knowledge-base of Contemporary Chinese (Yu et al. 2003), because these verbs cannot play as predicating words in most cases. As a result, we got 834 verbs in our collected data, which constitute the candidate verbs.

Then the task is designed to automatically determine the Up/Down properties of all the candidate verbs. Please note that the verb lexicon extraction in this paper is slightly different from previous literatures. Given two opposite sets of Up and Down seed verbs, our task is to determine the Up/Down properties of the verbs in the candidate verb set, rather than to expand the Up/Down lexicon.

4 Verb Lexicon Extraction

4.1 Using Tongyici Cilin

We extracted Up/Down seed words from Extended Tongyici Cilin (Cilin for short). Cilin is a manually built Chinese synonym thesaurus, which has been widely used in Chinese language processing. In Cilin there are five levels in the taxonomy structure, in which level 4 corresponds to synset.

In Cilin, synset of 涨/rise recommends Up while synset of 跌/drop recommends Down. So we extracted the synset of 涨/rise as Up words, and the synset of跌/drop as Down words. As a result, we got an Up lexicon (UpSeedSet) containing 34 up verbs and a Down lexicon (DownSeedSet) containing 36 down verbs. These verbs will serve as the seed words in the following verb lexicon extraction.

Then, the Up/Down trend of a candidate verb is defined as:

$$Trend(v) = \begin{cases} 1 \ if \ \ v \in Up \text{ Lexicon} \\ -1 \ if \ \ v \in Down \text{ Lexicon} \\ 0 \ \ if \ \ v \notin Up / Down \text{ Lexicon} \end{cases} \quad (1)$$

Also, we extracted the negation word set from Cilin, containing 25 words.

4.2 Using Web Search Engine

In order to verify our approach of similarity-based verb lexicon extraction, we conduct a contrast experiment as one baseline, following the work of Turney and Littman (2003), which has been widely applied in determining the semantic orientation of words in sentiment analysis. Adapting to our particular task, the Up/Down property of a verb is calculated from the strength of its association with a set of words indicating Up trend, minus the strength of its association with a set of words indicating Down trend. Accordingly, the Up/Down property of a verb is computed as:

$$SO(v) = \frac{\sum_{uv \in UpSeedSet} PMI(uv, v)}{|UpSeedSet|} - \frac{\sum_{dv \in DownSeedSet} PMI(dv, v)}{|DownSeedSet|} \quad (2)$$

Where *UpSeedSet*/*DownSeedSet* is the Up/Down seed word set extracted from *Cilin*. In this paper, we use *Baidu* search engine[1], as it is the most popular one in China.

Table 2 lists the top 5 examples in the extracted verb lexicon. That is, the top 5 verbs with the highest values of $SO(v)$ in the Up lexicon, and the top 5 verbs with the smallest values of $SO(v)$ in the Down lexicon.

Table 2. The top 5 examples in the extracted verb lexicon using web search engine

Up	簇拥/crowd, 撑腰/back up, 颤/quiver,伸手/stretch, 回望/recall
Down	开征/levy, 清仓/liquidate, 微升/rise slightly,保值/preserve, 看涨/rise

The Up/Down trend of a verb is defined as:

$$Trend(v) = \begin{cases} 1 \ if \ \ SO(v) > 0 \\ -1 \ if \ \ SO(v) < 0 \\ 0 \ \ if \ \ SO(v) = 0 \end{cases} \quad (3)$$

[1] http://www.baidu.com/

4.3 Using Chinese Semantic Similarity

Previous studies on Chinese lexical semantic similarity based on large corpora are quite limited, and this paper makes a study on this issue and proves its application usage in verb lexicon extraction.

Given a Chinese thesaurus of semantic similarity, the Up/Down trend of a candidate verb is computed using the following equation:

$$Trend(v) = \frac{\sum_{uv \in UpSeedSet} sim(uv, v)}{|UpSeedSet|} - \frac{\sum_{dv \in DownSeedSet} sim(dv, v)}{|DownSeedSet|} \quad (4)$$

Where $sim(w_1, w_2)$ denotes the similarity score between w_1 and w_2.

Lin's method (Lin, 1998) is adopted here to calculate Chinese lexical semantic similarity. The similarity $sim(w_1, w_2)$ between two words w_1 and w_2 is computed as follows:

$$\text{sim}(w_1, w_2) = \frac{\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T(w_2)} I(w_2, r, w)} \quad (5)$$

Where (w, r, w') is a dependency triple consisting of two words w, w' and the grammatical relationship r between them; $I(w, r, w')$ denotes the mutual information; $T(w)$ is the set of pairs (r, w') such that $I(w, r, w')$ is positive.

We use the corpus of Chinese Gigaword, provided by Linguistic Data Consortium (LDC). We take the part of Xinhua News Agency, stamped between the year of 1990 to 2004, containing 471,110K Chinese characters (1.6G) and totally 992,261 documents.

Considering the grammatical relationship *r* in Equation (5), two kinds of contexts are adopted: window-based and dependency-based.

Window-Based Method. All the texts in the corpus were automatically word-segmented and POS-tagged using the open software *ICTCLAS*[2]. The window context is defined as the left and right 3 words to the target word, along with their position offset to the target word.

Dependency-Based Method. All the texts in the corpus were automatically parsed using the Stanford Chinese dependency parser[3] (Chang et al., 2009). Then all the dependency triples were extracted. All of the 45 named grammatical relationships are regarded as the dependency contexts.

When calculating Chinese lexical semantic similarity, we only compute the similarity between two words sharing the same POS tag. In both window-based and dependency-based approaches, we generate up to 200 most similar words for each verb.

Table 3 and Table 4 list the top 5 examples in the extracted verb lexicon, by using window-based method and dependency-based method respectively.

---

[2] http://ictclas.org/

[3] http://nlp.stanford.edu/software/lex-parser.shtml

Table 3. The top 5 examples in the extracted verb lexicon using window-based Chinese semantic similarity

Up	提升/raise, 发展/develop, 奔/rush, 升级/increase, 跨入/stride
Down	升值/rise, 衰退/decline, 跌破/drop,上扬/rise, 抛售/undersell

Table 4. The top 5 examples in the extracted verb lexicon using dependency-based Chinese semantic similarity

Up	提升/raise, 掀起/lift, 活跃/enliven,攀/climb, 升温/warming
Down	跌破/drop, 接触/touch, 反弹/rebound,挫/ frustrate, 波动/fluctuate

4.4 Using English Semantic Similarity

4.4.1 Motivation

English lexical semantic similarity has been extensively studied, and has been applied to many NLP tasks. In this section, we try to apply English semantic similarity to Chinese verb lexicon Extraction.

Some studies on synonym extraction have exploited the cross-lingual knowledge, as discussed in Related Work. The assumptions of these methods as well as our methods are listed as below.

Assumption-Wu (Wu and Zhou, 2003)**:** Two words are synonyms if their translations are similar.

Assumption-Lin (Lin et al., 2003)**:** Translations of a word from another language are often synonyms of one another.

Assumption-Plas (Plas and Tiedemann, 2006)**:** Words sharing translational contexts are semantically related.

Our Assumption: If two words are semantically similar in a language, their translations in another language would be also similar with the same similarity score.

We extend the cross-lingual property from synonyms to similar words, where the former is only a small set of the latter. Instead of using translational contexts (Plas and Tiedemann, 2006), we manage to directly use the translational semantic similarity. Considering that lexical semantic similarity computation is time consuming with the computation cost $O(n^2m)$, where n is the number of target words and m is the number of context features, our approach provides a fast way to make use of the full-fledged English resources and techniques.

4.4.2 Methods

Using English similar words to extract Chinese verb lexicon includes the following two phases.

1) To automatically translate the seed words and candidate verbs into English;
2) To compute the Up/Down trend of each tr(v)∈*tr_CandidateVerbs* based on an English thesaurus of semantic similarity, using the following equation:

$$Trend(tr(v)) = \frac{\sum_{uv \in tr\_UpSeedSet} sim(uv, tr(v))}{|tr\_UpSeedSet|} - \frac{\sum_{dv \in tr\_DownSeedSet} sim(dv, tr(v))}{|tr\_DownSeedSet|} \quad (6)$$

Where $sim(w_1, w_2)$ denotes the similarity score between w_1 and w_2; $tr(v)$ is the translation of a candidate verb v; *tr_CandidateVerbs* are the translations of candidate verbs; *tr_UpSeedSet* and *tr_DownSeedSet* are the translations of Up seed words and Down seed words.

In the first phase, we use Google translator[4] to translate seed words and candidate verbs into English. In the second phase, we use Lin's proximity-based thesaurus of semantic similarity, which can be freely downloaded from DeKang Lin's homepage[5]. For each word, the thesaurus lists up to 200 most similar words and their similarity scores.

Then, the Up/Down trend of a candidate verb v is defined as that of the translation $tr(v)$: $Trend(v) = Trend(tr(v))$.

Table 5 lists the top 5 examples in the extracted verb lexicon using English semantic similarity, showing that the predicting results are promising.

Table 5. The top 5 examples in the extracted verb lexicon using English semantic similarity

Up	增/add, 加大/enlarge, 增长/increase, 增加/increase, 跳/jump
Down	崩/collapse, 倒塌/tumble, 崩溃/crash,撤退/retreat, 跌破/drop

4.5 Using Ensemble Methods

In order to further investigate the complementary property of English and Chinese semantic similarity in verb lexicon extraction, we conduct experiments using ensemble methods, by combining all of the three values computed under the methods of English semantic similarity($Td_{EN}(v)$), window-based Chinese semantic similarity ($Td_{Win}(v)$) and dependency-based Chinese semantic similarity ($Td_{Dep}(v)$).

Firstly, the Up/Down trend values in each method are normalized by dividing it with the max absolute value among all candidate verbs. For instance, the trend values with window-based Chinese semantic similarity method are normalized using the following Equation:

$$Td_{Win}(v) = \frac{Trend_{Win}(v)}{Max\{|Trend_{Win}(v^i)| \,\big|\, v^i \in CandidateVerbs\}} \quad (7)$$

We carry out the following four ensemble methods.

[4] http://translate.google.cn/

[5] http://webdocs.cs.ualberta.ca/~lindek/

Average. The new value is the average of all three values:

$$Trend^{Ens}(v) = \frac{Td_{EN}(v) + Td_{Win}(v) + Td_{Dep}(v)}{3} \quad (8)$$

Max. The new value is the one with the max absolute value among all three values:

$$Trend^{Ens}(v) = x, \ |x| = \max\{|Td_{EN}(v)|, |Td_{Win}(v)|, |Td_{Dep}(v)|\} \quad (9)$$

Min. The new value is the one with the min absolute value among all three values:

$$Trend^{Ens}(v) = x, \ |x| = \min\{|Td_{EN}(v)|, |Td_{Win}(v)|, |Td_{Dep}(v)|\} \quad (10)$$

Majority Voting. This combination result relies on the Up or Down polarity tags, rather than the absolute values. The polarity tag receiving more votes among three methods is chosen as the final tag.

5 Experiments

5.1 Data Collection

The data of blog titles was collected from *Sina* website. We manually picked out some blog titles from housing[6] and stock[7], written from January 1st to December 31th, 2009. The two authors of this paper manually labeled Up/Down/Unknown tag to each title in a doubly blind manner. The inter-annotator agreement is in a high level with a Kappa value of 0.81. Discarding the titles with Unknown tag, finally we picked out 1,000 titles for housing and 1000 titles for stock respectively, where titles tagged with Up and Down are evenly distributed. All the data was used as test data.

5.2 Experimental Results

Table 6 and Table 7 list the experimental results using 9 methods on two datasets, where the two methods of *Cilin* and *Web Search* serve as two baselines.

As is expected, the method using *Cilin*, which is manually complied, gets the highest precision but rather poor recall. But to our surprise, the *Web Search* method, which has been widely used for lexicon extraction in sentiment analysis, gets the worst performance in our task, and is even worse than the seed lexicon from *Cilin* according to F score.

Both the two methods of *English semantic similarity (English-Sim)* and *Window-based Chinese semantic similarity (Win-Chinese)* outperform two baselines substantially in F score, validating the effectiveness of applying similar words to verb lexicon extraction. But the performance of *Dependency-based Chinese semantic similarity (Dep-Chinese)* is quite poor, achieving an even lower F score than *Cilin* baseline in the housing market.

[6] http://bj.house.sina.com.cn/HouseBlog/

[7] http://blog.sina.com.cn/lm/114/111/day.html

By applying English similar words to Chinese verb lexicon extraction, we harvest competitive results in overall F score comparing with window-based Chinese similar words in both two datasets, validating our assumption of cross-lingual property of lexical semantic similarity.

Among ensemble methods, the *Average* and *Max* methods rival three individual approaches in both two datasets, proving the complementary property of three different semantic similarity scores. *Max* gets the best performance among 9 methods, which outperforms *Cilin* baseline by 13.51% in F score on housing data and 14.24% on stock data.

5.3 Discussion

Verb lexicon extraction using lexical semantic similarity improves the performance of market trend mining, especially in recall score. By applying verb extraction, many candidate verbs that do not appear in the seed words in *Cilin* are given an appropriate Up/Down property. There are 70 seed words in *Cilin*, and the number of indicator words is expanded to 309 using *English semantic similarity* method and 457 using *Window-based Chinese semantic similarity.* We guess that the poor performance of *Dependency-based Chinese semantic similarity* lies in the unsatisfactory performance of Chinese dependency parser.

Table 6. Experimental results on housing data[8]

Methods	UP(%)			DOWN(%)			OVERALL(%)		
	P	R	F	P	R	F	P	R	F
Cilin	**87.39**	38.80	53.74	**84.80**	34.60	49.15	**86.15**	36.70	51.47
Web Search	55.53	42.20	47.95	56.81	39.20	46.39	56.14	40.70	47.19
English-Sim	74.72	53.20	62.15	69.63	48.60	57.24	72.20	50.90	59.71
Dep-Chinese	60.89	27.40	37.79	51.54	63.40	56.86	54.05	45.40	49.35
Win-Chinese	72.06	45.40	55.71	61.29	**68.40**	64.65	65.18	56.90	60.76
Average	72.34	54.40	62.10	64.99	67.20	66.08	68.09	60.80	64.24
Max	72.85	**55.80**	**63.19**	65.88	67.20	**66.53**	68.87	**61.50**	**64.98**
Min	57.91	32.20	41.39	52.52	64.60	57.94	54.20	48.40	51.14
Voting	72.63	41.40	52.74	64.37	54.20	58.85	67.71	47.80	56.04

[8] P,R,F stand for Precision, Recall and F score.

Table 7. Experimental results on stock data

Methods	UP(%)			DOWN(%)			OVERALL(%)		
	P	R	F	P	R	F	P	R	F
Cilin	**88.81**	23.80	37.54	**85.71**	32.40	47.02	**87.00**	28.10	42.48
Web Search	50.28	35.80	41.82	48.19	37.20	41.99	49.19	36.50	41.91
English-Sim	71.28	**42.20**	**53.02**	64.80	48.60	55.54	67.66	45.40	54.34
Dep-Chinese	51.49	20.80	29.63	50.56	63.00	56.10	50.79	41.90	45.92
Win-Chinese	72.00	28.80	41.14	55.80	**70.20**	62.18	59.71	49.50	54.13
Average	66.08	37.80	48.09	57.12	67.40	61.83	60.05	52.60	56.08
Max	67.13	38.40	48.85	57.63	68.00	**62.39**	60.74	**53.20**	**56.72**
Min	63.83	36.00	46.04	55.89	66.40	60.69	58.45	51.20	54.58
Voting	64.61	23.00	33.92	56.40	58.20	57.28	58.50	40.60	47.93

The distributional hypothesis states that words sharing similar meanings tend to appear in similar contexts (Harris, 1968). In a strict view, distributional hypothesis generates words sharing semantic relatedness rather than semantic similarity. Some researchers believe that distributional related words minimize their usage in many applications, and only synonym words are useful and should be identified (e.g., Lin, 2003; Plas and Tiedemann, 2006). However, our experimental results show that semantically related words are of great use in verb lexicon extraction in mining market trend. Words like "跳/jump" "绽放/bloom" are assigned as Up trend while words like "滑/gliding" "倒塌/collapse" are assigned as Down trend. To our delight, some figurative meanings of words are also given correct Up/Down properties. For instance, "入冬/the beginning of winter", "降温/cooling down", "缩水/shrink" and "腰斩/cutting sb. in two at the waist" are given Down trend orientations. These words obviously are not the synonyms of 跌/drop, but are semantically related with 跌/drop. So, it is the more-general idea of semantic relatedness, rather than semantic similarity, that we need for some NLP applications.

To our delight, applying English similar words into Chinese verb lexicon extraction brings an obvious improvement in performance, which is well compared with *Window-based Chinese semantic similarity* method and far better than *Dependency-based Chinese semantic similarity* method. Considering the noises introduced by Google translator, applying English similar words would get better results. The experimental results confirm our assumption of cross-lingual property of semantic similarity. Considering the cost of lexical semantic similarity computation, our approach provides a fast way to make full use of English knowledge and resources in Chinese semantic similarity computation.

The poor performance using web search engine lies in two reasons. 1) The returned results of the Chinese search engine are not as good as English. 2) The noises

presented in the Chinese web are huger than English, mostly because that Chinese words are not naturally segmented. The experimental results tell us that some well-developed techniques using web search engine in English perhaps would not work well in Chinese.

6 Conclusion

In this paper, we make a thorough study on applying semantic similarity to verb lexicon extraction, aiming at the task of mine Up/Down market trend from blog titles. We harvest a great increase of F score by integrating English and Chinese semantic similarity using the *Max* ensemble method. Both the two methods of *English semantic similarity* and *Window-based Chinese semantic similarity* achieve promising results. Introducing English similar words to Chinese verb lexicon extraction brings an obvious improvement, which is well compared with the *Window-based Chinese semantic similarity* method but is far batter in computation cost.

Our experimental results show that verb lexicon extraction based on semantic similarity is of great use for some NLP applications. In future work, we will apply semantic similarity to other application usages. Also, our experimental results confirm the assumption of cross-lingual property of semantic similarity. In future work, we will make full use of English resources in Chinese semantic similarity computation.

We also realize that, the lexicon-based approach is far from enough to recognize the Up/Down trend expressed in blog titles, and there are still a lot of challenges in this task.

Acknowledgments. This work was supported by 2009 Chiang Ching-kuo Foundation for International Scholarly Exchange (under Grant No. RG013-D-09).

References

1. Chang, P., Tseng, H., Jurafsky, D., Christopher, D.: Discriminative reordering with Chinese grammatical relations features. In: Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation, pp. 51–59 (2009)
2. Chaudhuri, S., Ganti, V., Xin, D.: Exploiting web search to generate synonyms for entities. In: Proceedings of WWW 2009, pp. 166–172 (2009)
3. Chesley, P., Vincent, B., Xu, L., Srihari, R.: Using verbs and adjectives to automatically classify blog sentiment. In: Proceedings of AAAI 2006, pp. 27–29 (2006)
4. Peramunetilleke, D., Wong, R.: Currency exchange rate forecasting from news headlines. In: Proceedings of the 13th Australian Database Conference, pp. 129–137 (2002)
5. Harris, Z.: Mathematical Structures of Language. Wiley, New York (1968)
6. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of Coling/ACL 1998, pp. 768–774 (1998)
7. Lin, D., Zhao, S., Qin, L., And Zhou, M.: Identifying synonyms among distributional similar words. In: Proceedings of IJCAI 2003, pp. 1492–1493 (2003)
8. Liu, F., Wang, D., Li, B., Liu, Y.: Improving blog polarity classification via topic analysis and adaptive methods. In: Proceedings of NAACL 2010 (2003)

9. Potthast, M., Stein, B., Anderka, M.: A Wikipedia-based multilingual retrieval model. In: Proceedings of ECIR 2008 (2008)
10. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval (2008)
11. Pantel, P., Crestan, E., Borkovsky, A., Popescu, M., Vyas, V.: Web-scale distributional similarity and entity set expansion. In: Proceedings of EMNLP 2009 (2009)
12. Park, K., Jeong, Y., Myaeng, S.: Detecting experiences from weblogs. In: Proceedings of ACL 2010 (2010)
13. Pennacchiotti, M., Pantel, P.: Entity extension via ensemble semantics. In: Proceedings of EMNLP 2009 (2009)
14. Sorg, P., Cimiano, P.: Cross-lingual information retrieval with explicit semantic analysis. In: Proceedings of CLEF 2008 (2008)
15. Plas, L., Tiedemann, J.: Finding synonyms using automatic word alignment and measures of distributional similarity. In: Proceedings of COLING/ACL 2006, pp. 866–873 (2006)
16. Shi, S., Zhang, H., Yuan, X., Wen, J.: Corpus-based semantic class mining: distributional vs. pattern-based approaches. In: Proceedings of COLING 2010, pp. 993–1001 (2010)
17. Turney, P., Littman, M.: Measuring praise and criticism: inference of semantic orientation from association. ACM Transactions on Information Systems, 315–346 (2003)
18. Wu, H., Zhou, M.: Optimizing synonym extraction using monolingual and bilingual resources. In: Proceedings of the Second International Workshop on Paraphrasing: Paraphrase Acquisition and Application (2003)
19. Yu, S., Zhu, X., Wang, H., Zhang, H., Zhang, Y., Zhu, D., Lu, J., Guo, R.: The Grammatical Knowledge-base of Contemporary Chinese. Tsinghua University Press (2003)