Using normal dictionaries to extract multiple semantic relationships

Li Liang¹, Fang Tian²

¹Department of Computer Technology and Applications, Qinghai University, Xining, People’s Republic of China
²Information Technology Center of Qinghai University, Qinghai University, Xining, People’s Republic of China

E-mail: tianfang@qhu.edu.cn

Abstract: This article proposes a rule-based semantic relationship extraction method for Chinese sentiment lexicon to solve the problem that words in the sentiment dictionary cannot be updated in time, and expand the lexicon and the semantic relationship in the dictionary. Using the massive information on the Internet, the modern Chinese dictionary and sentiment dictionary and other resources, based on the interpretation of words and the characteristics of semantic relationships between words, sentiment lexicon are automatically extracted and their semantic relationships are annotated. At the same time, the semantic relationship extraction rules are defined, and the sentiment lexicon with synonymous relationship and antisense relationship is extracted recursively. The extraction rules and algorithms of emotional intensity relationship are also improved, and the lexicon is extracted according to the level of degree adverbs and their emotional weights are calculated. Compare the extraction effect before and after the improved algorithm and the change of the number of lexicon extracted after increasing the corpus. The experimental results show the effectiveness and accuracy of the proposed method.

1 Introduction

Relation extraction is designed to extract semantic relationships—such as synonyms or antonyms—between pairs of words from corpus data. In natural language processing, lexical semantic knowledge is becoming more and more important, especially in applications such as word sense disambiguation [1, 2], question answering [3, 4] and text summarisation [5]. Despite the wide coverage of existing hand-labelled resources such as WordNet [6], there are coverage issues for applications involving specific areas or applications other than English. With the development of the network, new words are constantly appearing, and the semantic relationship between lexicon is constantly generated. It is the labour and material to update the vocabulary by hand. In the face of the data-magnifying era of big data, massive network knowledge and dictionaries provide abundant resources for the acquisition of anti-sense relations and different emotional intensity relationships.

Semantic information was first proposed by Professor Bar-Hillel and Professor Carnap in 1953 [7]. The concept of semantic relations is clearly proposed by Professor John F. Sowa [8]. In 1992, he defined that the semantic network can be described as a directed graph with fixed points as vocabulary and edges as lexical semantic relationships. Semantics is the intrinsic meaning of linguistic form, which reflects people’s reflection and understanding of language. At present, there are some data resources in terms of synonyms and antonyms, such as TongYiCiCiLin [9] written by Mei Jiaju and others, Chinese Concept Dictionary, Hownet [10], Wordnet etc. Among them, Mei Jiaju and others compiled the TongYiCiCiLin in 1983. Since the vocabulary has not been updated for a long time, some words have become uncommon words, and many new words have not been added. Therefore, the information retrieval laboratory of Harbin Institute of Technology has invested a large amount of manpower and material resources to delete 14,706 non-common words and rare words in TongYiCiCiLin, and added a large number of related vocabulary to classify the original vocabulary. The above semantic dictionaries are constructed by hand, which has high maintenance cost, slow update speed and low vocabulary coverage.

There are several approaches to relation extraction. Pattern-based semantic relationship extraction has been widely used in research. Hurst [11] defines patterns to extract the underlying relationship, using a pattern such as ‘w1, such as 2’ to infer that w2 represents the lower word of w1, where w represents a noun phrase. Chkovski et al. [12] also used pattern to extract verb relationships based on large-scale corpus. They first collect verb pairs with strong relevance as candidates, and then formulate search queries by instantiating predefined patterns with verb pairs. The higher the frequency with which a verb appears in conjunction with a pattern, the more likely it is to have the relationship stated by the pattern. Lu Yong et al. [13] extracted the common patterns in which synonym appeared in the dictionary definition, and then extracted the synonyms according to the summarised model. These methods all rely on the pattern of human discovery and do not ensure that these patterns are comprehensive.

Other approaches to relation extraction use contextual information of single terms. NA Castro-sanchez et al. [14] use the upper and lower positional relationship in the semantic dictionary to obtain a synonym automatic acquisition method for a verb, which is limited to a verb in which the vocabulary in the definition and the defined verb have a top-bottom relationship. Hag iwara [15] measures context similarity through distribution features and uses support vector machines (SVM) to separate synonym pairs from other pairs. Mirkin et al.[16] used a pattern-based and distribution-based approach to collect term pairs as candidates and then trained the SVM classifier to maintain the correct pair. However, many distribution similarity calculations are computationally inefficient because they require time-consuming construction of dependency trees.

There are some other approaches. Lu Yong et al. [17] proposed to extract the synonym automatically based on dictionary definition. By collecting a large number of words containing the definition, construct the relationship diagram between vocabulary, establish the adjacency matrix and calculate the adjacency matrix using Pagerank algorithm. The work in [18] uses a dependency parse tree and defines some simple rules to extract relationships from free text. The above study extracts semantic relationships from the perspective of discovery. Although this method can extract semantic relations well, it has low coverage due to its dependence on existing resources. These shortcomings make the extended semantic dictionary have certain limitations.

The key difference in the approach we take in this paper is that the choice of corpus is authoritative and the rules we choose are comprehensive. The main contributions of our paper are summarised as follows:
(i) We use massive resources to extract the synonymous and antiseem relations from the emotional words in the existing sentiment dictionary. Improve the emotion intensity extraction algorithm to increase the number of extracted lexicon.

(ii) Based on the modifiers in front of the emotional lexicon, the emotional intensity of the extracted emotional lexicon is calculated. Refine the emotional level of the vocabulary.

The rest of this paper is structured as follows. Section 2 introduces the corpus used. In Section 3, we overview our approach. The experimental design, the datasets, the evaluation measures and the results are explained in Section 4. Conclusions and future works are in Section 5.

2 Normal dictionaries

In order to ensure the availability of the extracted emotional vocabulary, the selection of corpus should not only consider the validity and comprehensiveness of the content contained in it, but also consider the emotions contained therein. Based on this, this paper selects the authoritative modern Chinese dictionary as the basic corpus, and then adds the vocabulary crawled from Baidu and its interpretation to improve the coverage of the extracted emotional vocabulary.

The Modern Chinese Dictionary [19] was written by the Institute of Linguistics of the Chinese Academy of Social Sciences. This book basically covers the current stable use of Chinese words. The words and their explanations have strong authority, which guarantees the correctness of semantic relationship extraction. In modern Chinese dictionaries, each word has its corresponding interpretation and example sentences.

In order to make the coverage of the extracted emotional words more comprehensive, the corpus has added Baidu interpretation. It contains the annotated words in the Baidu Encyclopedia. Baidu Encyclopedia is an open, free online encyclopedia platform launched by Baidu [20]. The content of the body of the entry is structured according to the specificity. At the same time, in order to ensure the accuracy of the content, each content reference material is required to supplement the evidence. Therefore, crawling from Baidu Encyclopedia with comments is a word entry, and its accuracy and comprehensiveness are also guaranteed.

The content structure of Baidu entry and Dictionary is the same, merged into a common dictionary for semantic relationship extraction. The structure of their contents is shown in Fig. 1.

3 Methodology

Our approach consists mainly of two processes. The first is to clean the data in the normal dictionaries. The second is based on the relationship between lexicon and interpretation in the corpus, combined with the rule-based method for semantic relationship extraction. The extracted semantic relations mainly have synonymous relations, antiseem relations and emotional strength relationships.

3.1 Data preprocessing

The noise in the corpus can seriously affect the applicability of the extraction results, so it needs to be cleaned beforehand. The modern Chinese dictionary contains vocabulary, definitions and examples. Interpretation is the interpretation of vocabulary, which is highly relevant to vocabulary. However, the relationship between the example sentence and the vocabulary is not very close, which will affect the efficiency of the extraction and the correctness of the extracted results. Therefore, we should filter the example sentence before the experiment.

It has been observed that the example sentences in dictionary usually follow a colon and end with a period or question mark. Therefore, the definition rule 1 filters the example sentences in the modern Chinese dictionary. The specific rules are shown in Table 1. The knowledge captured from the Baidu Encyclopedia has no examples, so there is no need for clear data work.

\[
\text{Algorithm 1}
\]

\[
\begin{align*}
\text{Input: seed}_1 \\
\text{Output: Synonym for seed}_1 \\
\text{Step 1: word segmentation} \\
\text{Step 2: Cleaning data} \\
\text{Step 3: Add Seed}_1 \\
\text{Step 4: while(seed, in SA)} \\
\quad \text{if seed, not match Rule2} \\
\quad \quad \text{if seed, match Rule3} \\
\quad \quad \quad \text{Syn.append(seed)} \\
\text{Step 5: Syn update SA} \\
\text{Step 6: if(seed, not in SA), then step7; else return step4;} \\
\text{Step 7: end}
\end{align*}
\]

3.2 Extraction of synonyms in semantic relations

The interpretation of a word is the definition of it. If a word w1 appears in the interpretation of a word w2, then largely w2 is a synonym or for w1. How to accurately extract the required vocabulary needs to be implemented according to the corpus content structure and the corresponding extraction rules. Using emotional words as seeds, traversing the corpus to match the seeds to obtain lexicon with which they have synonymous relations.

Analysis of the corpus structure, there may be negative words before the seed, resulting in errors in the extraction results. Define rule 2 in Table 2 to filter out the antonyms first, and then extract the synonymous relationship of the seeds according to rule 3.

According to the defined rules, the recursive algorithm is used to extract the synonymous relationship of the emotional lexicon in the general dictionary. The emotional lexicon in the existing dictionary are input as seeds, and the input seeds are matched from the comments of the word l according to rule 3. If the match is successful, output is l. If the match is unsuccessful, continue traversing until the knowledge corpus is traversed. The implementation of the algorithm is as follows (Fig. 2).

3.3 Extraction of antonyms in semantic relations

According to the relationship between the interpretation and its defined words, if the lexicon l1 appears in the definition of a lexicon l2 and has a negative prefix modification, then largely l2 is the opposite of l1. The extraction of antonyms mainly considers the influence of negative words on emotional linguistic sentiment. As an experiment, the negative word dictionary constructed in this paper contains the negative words commonly used in Chinese language, which can extract the antiseem relationship more comprehensively. In addition, an antonym that has a synonymous relationship with a seed or a synonym that has an antiseem relationship with a seed can be extracted as an antonym of the seed.
As shown in the flow chart of the extraction method of the antonym of Fig. 3, it mainly includes the definition of the rule and the algorithm implementation.

The lexical extraction of the antisense relationship is carried out in two aspects: first, the synonym of the seed is extracted, and the seed Seed_s and its synonym are used together as the seed set seed_A to extract the antonym; the second is to extract the synonym from the first extracted antonym A1 as the seed seed_B. Rule 5 is defined by the relationship between corpus interpretation and emotional vocabulary. According to whether there is a negative word modification before the matching seed in the interpretation, whether the position is closer to the beginning word to extract the emotional vocabulary with anti-sense relationship. In order to ensure the correctness of the extraction result, the synonym extracted in the process should be closer to the seed. Therefore, the definition rule 4 extracts the synonym with the same emotional intensity of the seed. The specific rules are shown in Table 3.

According to the defined rules, the algorithm 2 is written to extract the antisense relationship from the ordinary dictionary. Use the vocabulary of the existing sentiment dictionary as input and match it with the vocabulary l annotation in the dictionary. If the match is successful and rule 4 is satisfied, then l is output. If it does not match, continue to traverse the dictionary until the dictionary is traversed. The implementation of the algorithm is as follows (Fig. 4).

3.4 Lexicon extraction with different emotional intensity

Degree adverbs have a great influence on the emotional weight of vocabulary. Different levels of adverbs have different effects on the intensity of emotional words. As shown in the following example, the three-sentence emotion increases step by step.

I'm sad.
I am very sad.
I am extremely sad.

As shown in the flow chart of the extraction method of the antonym of Fig. 3, it mainly includes the definition of the rule and the algorithm implementation.

The lexical extraction of the antisense relationship is carried out in two aspects: first, the synonym of the seed is extracted, and the seed Seed_s and its synonym are used together as the seed set seed_A to extract the antonym; the second is to extract the synonym from the first extracted antonym A1 as the seed seed_B. Rule 5 is defined by the relationship between corpus interpretation and emotional vocabulary. According to whether there is a negative word modification before the matching seed in the interpretation, whether the position is closer to the beginning word to extract the emotional vocabulary with anti-sense relationship. In order to ensure the correctness of the extraction result, the synonym extracted in the process should be closer to the seed. Therefore, the definition rule 4 extracts the synonym with the same emotional intensity of the seed. The specific rules are shown in Table 3.

According to the defined rules, the algorithm 2 is written to extract the antisense relationship from the ordinary dictionary. Use the vocabulary of the existing sentiment dictionary as input and match it with the vocabulary l annotation in the dictionary. If the match is successful and rule 4 is satisfied, then l is output. If it does not match, continue to traverse the dictionary until the dictionary is traversed. The implementation of the algorithm is as follows (Fig. 4).

3.4 Lexicon extraction with different emotional intensity

Degree adverbs have a great influence on the emotional weight of vocabulary. Different levels of adverbs have different effects on the intensity of emotional words. As shown in the following example, the three-sentence emotion increases step by step.

I'm sad.
I am very sad.
I am extremely sad.
3.5 Emotional intensity calculation

The emotional lexicon has different modifiers before and after, and the corresponding emotional intensity is different. The emotional intensity is divided according to the pre-seed modifier. The level of lexicon sentiment intensity is divided as shown in Table 6.

If the seed $s$ in the interpretation has a degree adverb modification, the emotional weight of the defined vocabulary $w$ will also change, and the emotional weight calculation of $w$ is defined as formula (1).

$$w = v + s$$
$$W_{\text{Emotion}}(w) = G(v) \ast W_{\text{Emotion}}(s)$$

The meaning of the letters in the formula:

- $w$: The interpreted words in the knowledge base
- $v$: degree adverb
- $s$: The words matched by the seeds in the interpretation
- $G(v)$: $v$ intensity value

If the seed $s$ in the interpretation has a negative prefix, the calculation of the emotional weight corresponding to $w$ is defined as formula (2).

$$W_{\text{Emotion}}(w) = (-1)^{u} \sqrt{W_{\text{Emotion}}(s)}$$

The meaning of the letters in the formula:

- $w$: Number of negative words before seed

4.1 Experimental data

In data processing, we define rules, filter the example sentences in the modern Chinese dictionary and process the word segmentation to generate preliminary experimental corpus.

The corpus used in the extraction of emotional lexical semantic relations is the modern Chinese dictionary and Baidu interpretation. The corpus is in text format, and each knowledge contains vocabulary and its definition. Experiment 1 compares the changes in the number of vocabulary extracted after adding Baidu entries to test whether the performance of our method is improved when the basic corpus is added. Experiment 2 was compared with previous studies to test whether the improved algorithm for emotional intensity is effective. Experiment 3 evaluated the correct rate of the method used in this paper.

4.2 Experimental results and analysis

The number of emotional words extracted from the modern Chinese dictionary is limited. In order to extract as many emotional words as possible, the corpus is added to the Baidu entry. The experiment compares the changes in the number of words extracted before and after adding Baidu entries. We randomly select three seeds $s_1$, $s_2$, $s_3$, and extract their synonymous relationship, antonym relationship and emotional intensity relationship. Figs. 7–9 are comparisons of the quantities of these three relationships after adding new corpus. The experimental results show that the addition
of Baidu interpretation has greatly increased the number of lexicon extracted.

Compared with the previous research, this paper considers synonymous words with the same intensity as seeds as the second kind of seeds, and extracts emotional words with different emotional strength. The graph a in Fig. 10 is the extraction method we have studied previously, and the graph b is the extraction method for the experiment in this paper.

The experiment randomly extracts three emotional words T1, T2, and T3 as emotion seeds, and extracts seeds by using the rules and algorithms defined in this paper. The seed extraction results are shown in Fig. 11. It can be seen that compared with the previous experiment, the number of extracted words is significantly increased, and the expansion of the original sentiment dictionary can be realised.

In order to verify the precision of this paper, the evaluation criteria use accuracy to evaluate performance. According to the characteristics of the algorithm, this paper defines the precision.

\[
\text{Precision} = \frac{\text{Extract the correct number of words}}{\text{Number of words extracted}}
\]

We randomly selected 10 emotional words as seeds to extract vocabulary with synonymous, antisense or emotional strength relationships, and the results were verified manually. The vocabulary accuracy of the three relationships in the results is shown in Table 7. The experimental results verify the effectiveness of the method.

5 Conclusion
This paper proposes a multi-semantic relation extraction algorithm based on general dictionary, which uses rule matching method to automatically extract emotional lexical semantic relationship. It solves the problem of manual labelling time and labour cost. This method does not rely on the existing sentiment dictionary, so it can recognise the emotional vocabulary beyond the existing sentiment lexicon, and can meet the requirements of expanding the dictionary or adding new semantic relations. After experimental verification, using the method proposed in this paper, the correctness of the extraction results is higher and the availability is achieved. Extracting vocabulary can also effectively expand the sentiment dictionary. The attempts made in this paper will likely be applied to the identification and lexical extraction of semantic relations of entities outside the emotional domain.

6 Acknowledgments
The work is supported by the foundational application research of Qinghai Province Science and Technology Fund (No. 2016-ZJ-743).

Table 7 Experimental result

|                   | Synonymous relationship | Antisense relationship | Emotional intensity |
|-------------------|-------------------------|------------------------|---------------------|
| precision (%)     | 96.25                   | 97.43                  | 96.22               |

7 References
[1] Zhang, J., Deng, B., Li, X.: ‘Concept based query expansion using wordnet’. Proc. of the 2009 Int. e-Conf. on advanced science and technology, Dajeon, Republic of Korea, 2009, pp. 52–55
[2] Ananthanarayanan, R., Chenthamarakshan, V., Deshpande, P.M., et al.: ‘Rule based synonyms for entity extraction from noisy text’. Proc. of the second workshop on Analytics for noisy unstructured text data, Singapore, 2008, pp. 31–38
[3] Van Der Plas, L., Bouma, G.: ‘Automatic acquisition of lexico-semantic knowledge for QA’. Proc. of OntoLex 2005- Ontologies and Lexical Resources, Jeju Island, Republic of Korea, 2005
[4] Oramas, S., Sordo, M., Espinosa-Anke, L.: ‘A rule-based approach to extracting relations from music tidbits’. Proc. of the 24th Int. Conf. on World Wide Web, Florence, Italy, 2015, pp. 661–666
[5] Yap, W., Baldwin, T.: ‘Experiments on pattern-based relation learning’. Proc. of the 18th ACM Conf. on Information and knowledge management, Hong Kong, People’s Republic of China, 2009, pp. 1657–1660
[6] Miller, G.A.: ‘Wordnet: a lexical database for English’, Commun. ACM, 1995, 38(11), pp. 39–41
[7] Bar-Hillel, Y., Carp, R.: ‘Semantic information’, Br. J. Philos. Sci., 1953, 4, (14), pp. 147–157
[8] Sowa, J.F.: ‘Conceptual graphs as a universal knowledge representation’, Comput. Math. Appl., 1992, 23(2–5), pp. 75–93
[9] Jiaju, M.: ‘TongyiciCilin’ (Shanghai Dictionary Press, Shanghai, People’s Republic of China, 1983)
[10] Zhendong, D.: ‘HowNet’, available at http://www.Keenage.com
[11] Hearst, M.A.: ‘Automatic acquisition of hyponyms from large text corpora’. Proc. of the 14th Conf. on Computational linguistics-Volume 2, Nantes, France, 1992, pp. 539–545
[12] Pantel, T., Verboeckhoven, P.: ‘Mining the web for fine-grained semantic verb relations’. Proc. of the 2004 Conf. on Empirical Methods in Natural Language Processing, Barcelona, Spain, 2004
[13] Yong, L., Hanqing, H.: 'Automatic extraction of Chinese synonyms based on pattern matching', J. Inf., 2007, 25, (6), pp. 720–724
[14] Castro-Sánchez, N.A., Sidorov, G.: 'Automatic acquisition of synonyms of verbs from an explanatory dictionary using hyponym and hyperonym relations'. Mexican Conf. on Pattern Recognition, Berlin, Heidelberg, 2011, pp. 322–331
[15] Hagiwara, M.: 'A supervised learning approach to automatic synonym identification based on distributional features'. Proc. of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop, Columbus, OH, USA., 2008, pp. 1–6
[16] Mirkin, S., Dagan, I., Geffet, M.: 'Integrating pattern-based and distributional similarity methods for lexical entailment acquisition'. Proc. of the COLING/ACL on Main Conf. poster sessions, Sydney, Australia, 2006, pp. 579–586
[17] Yong, L., Hanqing, H.: 'Automatic extraction of Chinese synonyms based on pagerank algorithm', J. Xihua Univ. Natural Sci. Edn., 2008, 27, (2), pp. 13–15
[18] Fundel, K., Kühn, R., Zimmer, R.: 'Relex—relation extraction using dependency parse trees', Bioinf., 2006, 23, (3), pp. 365–371
[19] Dictionary of the Institute of Linguistics, Chinese Academy of Social Sciences. Modern Chinese Dictionary. 5th Edition [M], The Commercial Press, 2005
[20] Zhao, J.: 'Construction of large-scale Chinese semantic dictionary', (Harbin Institute of Technology, People's Republic of China, 2011)
[21] Liang, L., Tian, F., Ren, F.: 'Automatic acquiring of Chinese semantic relations based on knowledge corpus'. 2018 IEEE Int. Conf. of Safety Produce Informatization (IICSPI), Chongqing, People's Republic of China, 2019, pp. 371–377