Contextual Information Based Technical Synonym Extraction

Yuxin WANG, Nobuyuki SHIMIZU, Minoru YOSHIDA, and Hiroshi NAKAGAWA
The University of Tokyo
{mini_wang,shimizu,minoru,nakagawa}@r.dl.itc.u-tokyo.ac.jp

ABSTRACT

Popular methods for acquiring synonymous word pairs from a corpus usually require a similarity metric based on contextual information between two words, such as cosine similarity. This metric enables us to retrieve words similar to a query word, and we identify true synonyms from the list of synonym candidates. Instead of stopping at this point, we propose to go further by analyzing word similarity network induced by the similarity metric and re-ranking the synonym candidates -- a mutual re-ranking method (MRM). We apply our method to a specific domain: technical synonym extraction from aviation reports in Japanese. Even though the Technical Corpus is small and the contextual information is sparse, the experimental result shows the effectiveness of applying the contextual information on extracting technical synonyms in Japanese, and that MRM boosts the quality of acquired synonyms.

1. Introduction

Without a good domain-specific thesaurus, we may grossly over or underestimate counts of important events in a corpus of incident reports, for example. Especially for the domain in which we are interested, it is not easy to automatically acquire synonyms. Technical incident reports in Japanese mixes English expressions, equivalent Japanese translation in both Kanji (Chinese) characters and Hiragana/Katakana (phonetic) characters, and their various abbreviations, making synonym acquisition quite challenging.

Various methods have been proposed for synonym acquisition. Among them, the most popular methods are based on distributional hypothesis (Harris, 1985): it states that synonym nouns share the similar contextual information. In synonym acquisition, this hypothesis is generally implemented as follows. In the first step, from a corpus to extract the statistics on contextual features of each word that are deemed important, and then each word is represented by a vector of these contextual features. In the second step, to choose a similarity metric such as cosine similarity and apply it to pairs of query words and synonym candidates, producing ranked lists of synonym candidates ordered by their similarity scores. Finally, to select top candidates from a ranked list and they are seen as synonymous with the query word.

Our proposal is to add a third step after the second. By examining the word similarity network induced by the similarity metric in the second step, we obtained a mutual re-ranking method (MRM) that takes accounts of the structure of the network. Since the network exhibits a scale-free property, we treat a hub word and non-hub word differently in re-ranking the ranked lists. While synonym relation is symmetric, our MRM is not. To the best of the authors' knowledge, no prior work uses structural information of the word similarity network.

We continue with other prior work of note. Following the distributional hypothesis, Hagiwara et al. (2006) examine the selection of useful contextual features. They compare the contributions of different types of contextual features using three general corpora in English, including dependency relations (subjects and objects of verbs, and modifiers of nouns) and proximity. They show that among them modification category has the greatest contribution and the combination of all types of contextual features perform the best in English. Terada et al. (2006) automatically acquire synonyms from technical incident reports in Japanese, using proximity features. They experiment a window size 0~4 (number of proceeding and succeeding terms), and the results show that the window size 2 is most effective for technical synonym acquisition in Japanese. They have not studied contributions of other types of contextual features in synonym acquisition.

Another objective of this paper is to evaluate contextual features similar to (Hagiwara et al., 2006) in acquiring synonyms from a corpus of technical aviation reports in Japanese. We experimentally investigate the effectiveness of different contextual features, and show how contributions of the contextual features differ from English to Japanese.

The rest of the paper is organized as follows. First, we describe the contextual information that are applied to technical synonym extraction in Japanese in Section 2; then in Section 3 we introduce the overall method: cosine for similarity calculation, MRM for synonym candidate ranking, automatically evaluation by handcrafted thesaurus. We report the experiment results with related discussions in Section 4. Finally, Section 5 concludes this paper.

2. Contextual Information

In this study we focus on acquiring synonymous nouns while the similar framework is applicable to other categories of words. From the set of nouns in the corpus, we disregard nouns of which term frequency is less than a predetermined threshold. We then extract the contextual features of the target nouns. As the morphological analysis structure of Japanese is different from English, re-implementation of (Hagiwara et al., 2006) is not feasible. Instead, we focused on the contextual features of child and parent constituents, and proximity in response with that on subjects and objects of verbs, and proximity in English. The contextual features of modifiers of nouns in English is not be considered, since they are mostly overlapping with the contextual features of child constituents in the output of Cabocha, the Japanese morphological analysis tool we used in the experiments for the study in this paper.

We extract three types of contextual features for technical synonym extraction in Japanese: dependency relations including child and dependency constituents, and proximity of neighboring words within a window size.

2.1 Dependency Relations

The first two types of contextual features we utilized are dependency relations, the child and parent constituents.

The dependency relations generally mean the predicate-argument structure, and for English, this includes subjects and objects of verbs, and modifiers of nouns. For Japanese, although Cabocha is a popular Dependency Structure Analyzer, capable of tokenizing a sentence and displaying the POS tags, dependency relations between tokens, it is incapable of finding the subjects or objects of verbs.

Since the extracted dependency structure depends on the analyzing tool and language, to absorb the differences between English and Japanese we turn to utilize child and parent constituents among tokens in Japanese instead of subjects and objects of verbs in English.

From the sentence chunks output from Cabocha, if a chunk contains a target noun, other chunks that are linked from this chunk via a dependency relation is treated as the child features of the target noun; similarly, if there are chunks that link to the chunk with the target noun via a dependency relation, this is treated as the parent features of the target noun.

Since we suppose the punctuation marks and partial words will not contribute to technical synonym acquisition in Japanese, we omit them from contextual features we consider, i.e. both the punctuation marks and partial words are filtered from the set of features.

For each target noun, we count the term frequency of every feature on both child and parent constituents and use them as components of a vector representing the target noun. We use a similarity metric to measure the similarity between two vectors, which is described in subsection 3.1.

2.2 Proximity

The third kind of contextual features we utilized is proximity. The neighboring words (proceeding and succeeding words) of the target noun in the same sentence are used as features. These features are justified on the basic assumption that two words with the similar meaning always share the similar distribution of proceeding and succeeding words. Its effectiveness on synonym acquisition has been shown not only in English (Baroni and Bisi, 2004) but also in Japanese (Terada et al. 2006). Since the reported window sizes of useful proximity features in English and Japanese are different, we will test different window sizes in the study.

Again, we do not use the punctuation marks and partial words as features. For each target noun, we count the term frequency of words that appear each side of the target noun and use them as components of a vector representing the target noun.

3. Synonym Extraction Method

3.1 Similarity Metric

We adopt vector space model (VSM) for calculating the similarity of target nouns in our study. Although VSM is simple, it is a popular formalism and its effectiveness has been shown in synonym acquisition (Terada et al., 2006; Hagiwara et al., 2006). Each target noun is represented as a vector; the dimensions of the vector represent contextual features, and their values are the weighted frequency of features in the context of the target word. We define *tf* as the term frequency of each feature word corresponding to the contextual features.

To calculate the similarity of two target nouns, Hagiwara et al. (2006) employ *tf·idf* weighting scheme for synonym acquisition in English, while Terada et al. (2006) has shown that for synonym acquisition in Japanese, *log(tf+1)* weighting scheme is more effective than that of using *tf·idf*. In this study we adopted the *log(tf+1)* weighting scheme.

Consequently, the similarity between two target nouns *x* and *y* is calculated as the cosine value of two corresponding vectors as below,

$$\cos(x, y) = \frac{\sum_{1\le i\le n} x_i y_i}{\sqrt{\sum_{1\le i\le n} x_i^2}\sqrt{\sum_{1\le i\le n} y_i^2}} \qquad (1)$$

Where, *n* is the total number of the feature words and x_i is the feature value of the *i*-th feature word, i.e. *log(tf+1)*.

3.2 Mutual Ranking Method

Suppose we have a query word *x* and a synonym candidate *y*. If *x* is truly similar to *y*, we can safely assume that the reverse is true: *y* would be similar to *x*, treating *y* as a query word and *x* as a synonym candidate. Based on this idea, we propose the mutual re-ranking method (MRM).

Suppose again that we have a query word *x*. Then using a similarity metric, we obtain a ranked list of *x*'s synonym candidates in ascending order. We will re-rank the synonym candidates in this original list using the rank score (*RS*). First, from the original list, we pick a synonym candidate *y*. Let us call the rank of *y* in the original list *rk(x,y)*. Treating *y* as a query, we then obtain a ranked list of *y*'s synonym candidates. In this list, *x* appears at some point. Let us call the rank of *x* in *y*'s ranked list *rk(y,x)*.

In other words, among edges adjacent to *x*, *rk(x,y)*-th heaviest one connects *x* to *y*. Vice versa is true for *rk(y,x)*. We define the rank score (*RS*) as follows:

$$RS(x, y) = A\log(rk(x, y)) + \log(rk(y, x)) \qquad (2)$$

Where, *A* is the coefficient for combine the mutual ranks between *x* and *y*. We then calculate *RS(x,y)* for every *y* in the original ranked list of x's synonym candidates, and re-rank them by RS in ascending order.

3.3 Evaluation Measure

To automatically evaluate the performance of the technical synonym acquisition method, we use a handcrafted thesaurus because many of the technical terms are not registered in a general dictionary. We compare all the terms in the corpus with a frequency above a predetermined threshold against the entries in the thesaurus.

Synonym candidates are prepared as follows: after all the similarities of each pair of target nouns are calculated, for every target noun all the other target nouns are going to be ranked based on the similarity in descending order, and then the ranked list is treated as the synonym candidates for the corresponding target noun.

To measure the performance, we use average precision. The precision of the synonym candidate noun which is ranked the k-th for the current target noun is given by,

$$precision\ (k) = \frac{1}{k}\sum_{1\le i\le k} r_i \qquad (3)$$

Where, r_k=1 if it is a correct synonym, r_k=0 otherwise. Consequently, the average precision of all the terms in the synonym candidate list for the current target noun is given by,

$$Ave\,\mathrm{Pr}\,e = \frac{1}{|\,D_q\,|}\sum_{1\le j\le N}(r_j \times precision(j)) \qquad (4)$$

Where, N is the number of all the synonym candidates, D_q is the number of correct synonyms.

4. Experiment

In our experiments, we use JAL-FDM report (4.5MB) as the Corpus and the Japanese Dependency Structure Analyzer, Cabocha. The POS tags in the output of Cabocha[1] is based on Chasen[2] are used to distinguish nouns. Terms with noun POS tag, unknown tag (typically terms in English), and symbol tag (again terms in English) except number (labeled as noun-number) and punctuation mark (labeled as symbol-general) are treated as nouns.

4.1 Usefulness of Contextual Features

Our experiments evaluate effectiveness of three types of contextual features: child (dep1 hereafter) and parent (dep2 hereafter) constituents, and proximity respectively. For proximity features, we test window size from 1 to 3 (represented as prox1, prox2, and prox3 respectively hereafter) to compare the best window size to capture the most useful contextual features. The experimental results are shown in Figure 1 with the legend of "no MRM".

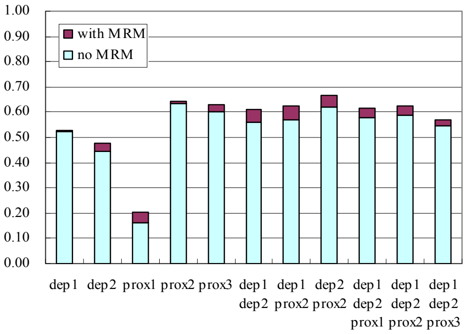

Figure 1. Performance with various contextual features and their combinations in addition to MRM.

The results show that while using the three types of contextual features for synonym extraction in Japanese, although child and parent constituents performed relatively well, proximity features in window size 2 performed the best (64.3%). Proximity features with window size 1 performed the worst, about 16.2% only, indicating that small window size does not provide enough useful contextual features for synonym acquisition in Japanese. On the other hand, bigger window does not mean more useful features are provided since proximity features with window size 3 performed not better than that with window size 2 at all.

Since proximity features with window size 2 performed the best, we conduct more experiments on combinations of every two types of the contextual features of dep1, dep2, and prox2. The results are shown in Figure 1 in the middle. The experimental results show that even though both dep1 and dep2 performed no better than prox2, the three combinations with any two features performed about the same; however, neither of the combination outperform prox2.

Next we combine the three types of contextual features together and vary the proximity window sizes from 1 to 3. The results are shown together in Figure 1 in the right hand. The experimental results show that the combination of dep1, dep2, and prox2 performed the best again. While this is not a surprising, we notice that the combination of dep1, dep2, and prox1 performed only 1% lower than that of the combination of dep1, dep2, and prox2. This is better than that of the combination of dep1, dep2, and prox3. We postulate that this indicates larger window of proximity provides more features which overlap with child and parent features. Although window size 1 is too small to provide enough useful features by itself (only 16.2%), the combination of dep1, dep2 and prox1 provides enough information leading to a stable performance. We think this is because features provided by window size 1 do not overlap with dep1 and dep2 very much.

However, feature combinations that use all types of contextual features are lower than using proximity contextual features with window size 2 alone. This fact implied that the importance of contextual features differs and different weights should be applied to each contextual

[1] Cabocha: Yet Another Japanese Dependency Structure Analyzer. http://chasen.org/~taku/software/cabocha/.

[2] Chasen: Japanese Morphological Analyzer. http://chasen.naist.jp/hiki/ChaSen/.

feature when combining them together for technical synonym acquisition in Japanese.

4.2 Importance of Contextual Features

Our next experiments extensively examine combinations of feature types with various weights. The experiments are done for the best performing combinations of feature types, i.e. the combination of dep1, dep2, and prox2.

The result shows when the weights to combine dep1, dep2, and prox2 are 0.25, 0.5, and 5 respectively, the performance of the combination reaches the best (64.3%). This finally outperforms simply using prox2 (63.4%). Even though in this case, the improvement in the average precision is relatively limited (0.9% only), it confirms the contextual features of parent and child constituents are useful, since the average precision is very difficult to improve when the base performance is already high. In addition, the lower weights on dep1 and dep2 do not mean dep1 and dep2 are nearly useless (their contributions are already shown in Figure 1). We think this indicates that there are relatively more unique characteristic features in prox2 than that in dep1 and dep2. We conclude that all the three types of contextual features utilized in the study are quite useful for technical synonym acquisition in Japanese.

4.3 Effectiveness of MRM

Our experiments of MRM with various coefficients *A* with different contextual features and their combinations, the results are shown in Figure 1 with the legend of "with MRM". For the combinations of three kinds of contextual features, the results show that when the coefficient is 1/2, the performance of synonym acquisition peaks for almost all combinations. In the case of the combination dep1, dep2, and prox2, the optimal setting for *A* improves the performance of synonym acquisition from 64.3% to 66.1%.

We plot the degrees of nodes in the word similarity network, counting only the edges heavier than threshold 0.39. This is shown in Figure 2. Since the network exhibits a scale-free property, it is reasonable to treat a hub word and non-hub word differently in re-ranking the ranked lists.

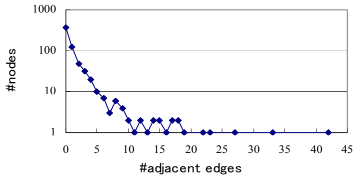

Figure 2. Word similarity network.

We checked the re-ranked synonym candidate list and compared it with that applied MRM before, and found that MRM can make the real synonym-like words' rank increased. For an example, the correct synonym of current target noun "オーダー" (katakana of order) is "order", before applying MRM, the rank of it is 2, after applying MRM, the rank increased to 1. Of course, MRM can not assure to increase the ranks of all the correct synonyms, but the experimental results show that MRM can make the overall performance improved.

5. CONCLUSION

In this paper, we propose a method for re-ranking the synonym candidates -- a mutual re-ranking method (MRM) through word similarity network. We apply the method to a specific domain: technical synonym acquisition based on contextual information from aviation reports in Japanese.

Our contributions are two folds. First we propose the mutual re-ranking method to improve the ranked list of synonym candidates for which the word similarity network exhibits a scale-free property. The extensive experiments on various setting show the effectiveness of the method, sometimes improving as much as 1.8%. The second contribution is investigation of utilizing three types of contextual features: child and parent constituents and proximity, for the technical synonym extraction. Even though our corpus of aviation incident reports is small and the contextual features are sparse, the experimental results show the effectiveness of utilizing the three types of contextual features on technical synonym acquisition.

Through the handcrafted thesaurus for evaluation, the experimental results show that among the three types of contextual features, child and parent constituents contribute relatively well, while proximity features with window size 2 contributes the best. When combining the three types of contextual features together, the case when the largest weight was put on proximity features performed the best, showing the importance of proximity contextual features on technical synonym acquisition.

Technical synonym acquisition in Japanese is extremely useful in various natural language applications, especially on text analysis and text mining. As the future work, we plan to investigate other types of contextual features, such as the relationship with sentence structure related particles, aiming to further improve the average precision on technical synonym acquisition.

6. REFERENCES

[1] Zellig Harris, 1985. Distributional Structure. *The Philosophy of Linguistics*. Oxford University Press, pages 26-47.

[2] Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiro Toyama. 2006. Selection of Effective Contextual Information for Automatic Synonym Acquisition. *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 353-360.

[3] Akira Terada, Minoru Yoshida, and Hiroshi Nakagawa. 2006. A Tool for Constructing a Synonym Dictionary using Context Information. *IPSJ SIG Technical Report, 2006-NL-176*, pages 87-94.

[4] Macro Baroni and Sabrina Bisi. 2004. Using cooccurrence statistics and the web to discover synonyms in a technical language. *Proc. of the Fourth International Conference on Language Resources and Evaluation (LREC 2004).*