Synonym Acquisition Using Bilingual Comparable Corpora

Daniel Andrade **Masaaki Tsuchida** **Takashi Onishi** **Kai Ishikawa**
Knowledge Discovery Research Laboratories, NEC Corporation, Nara, Japan
{s-andrade@cj, m-tsuchida@cq,
t-onishi@bq, k-ishikawa@dq}.jp.nec.com

Abstract

Various successful methods for synonym acquisition are based on comparing context vectors acquired from a monolingual corpus. However, a domain-specific corpus might be limited in size and, as a consequence, a query term's context vector can be sparse. Furthermore, even terms in a domain-specific corpus are sometimes ambiguous, which makes it desirable to be able to find the synonyms related to only one word sense. We introduce a new method for enriching a query term's context vector by using the context vectors of a query term's translations which are extracted from a comparable corpus. Our experimental evaluation shows, that the proposed method can improve synonym acquisition. Furthermore, by selecting appropriate translations, the user is able to prime the query term to one sense.

1 Introduction

Acquiring synonyms or near synonyms is important for various applications in NLP, like, for example, paraphrasing and recognizing textual entailment (Bentivogli et al., 2009).

For these tasks, lexical resources like WordNet are often used to improve performance. Although these resources provide good coverage in general domains, they lack vocabulary specific to certain domains. Other problems are the limited availability and size of lexical resources for languages other than English. [1]

As a consequence, various previous works (Grefenstette, 1994) among others, suggest to acquire synonyms and other semantically related words automatically from a monolingual corpus. The key assumption is that semantically similar words occur in similar context.

In general the larger the size of the monolingual corpus, the better and more detailed, we can extract the context, or context vectors for each relevant word (Curran and Moens, 2002). However, in a specific domain, the given monolingual corpus might be limited to a small size which leads to sparse context vectors. Another problem is that even for a specific domain, words can be ambiguous, which makes it unclear for which sense we are searching a synonym. For example, in the automobile domain, the ambiguous Japanese word バルブ (bulb, valve) has the synonyms 電球 (bulb) or 弁 (valve), depending on the meaning intended by the user.[2]

Our work tries to overcome both of these problems by enriching a context vector of a query word using *the context vectors of its translations* obtained from a comparable corpus in a different language. This way, some of the zero entries of a sparse context vector can be filled, and also disambiguation of the query word is possible. For example, if the desired query word is バルブ (bulb, valve), the user can select the auxiliary translation ”valve” in order to mark the ”valve” sense of the query word. Then, our system enforces the common parts of the context vector of バルブ (bulb, valve) and the context vector of ”valve”. Subsequently, when comparing the resulting context vector to synonym candidates' context vectors, the synonym 弁 (valve) will get a higher similarity score than the synonym 電球 (bulb).

In two experiments, we compare the proposed method to the baseline method which uses only the context vector obtained from the monolingual corpus. In the first experiment, the proposed method

[1] For example, the coverage of words for the English WordNet is 147,278 words, whereas for Japanese WordNet's coverage is only 93,834 words.

[2] To show the English meaning of a Japanese word, we place the English translations in brackets, directly after the Japanese word.

International Joint Conference on Natural Language Processing, pages 1077–1081,
Nagoya, Japan, 14-18 October 2013.

use all translations of a query term. In the second experiment, we use only the translations related to a certain sense of the query term. In both experiments the proposed method outperforms the baseline method, which suggests that our method is able to overcome sparsity and ambiguity problems.

In the following section we briefly embed our work into other related work. In Section 3, we explain our method in detail, followed by the two empirical evaluations in Section 4. We summarize our contributions in Section 5.

2 Related Work

Most work on synonym acquisition like (Grefenstette, 1994; Curran and Moens, 2002; Weeds and Weir, 2005; Kazama et al., 2010; Lin, 1998), contains basically of two steps: context vector extraction, and context vector comparison. In the first step, for the query term and each synonym candidate a context vector is extracted from the monolingual corpus. The context vector contains for example in each dimension how often the word co-occurred with another word in a certain syntactic dependency position. In the second step the query term's context vector is compared with each synonym candidate's context vector, for example by using the cosine similarity.

The problem of sparse context vectors, i.e. many dimensions in the context vector which contain zero entries, can be addressed by truncated Singular Value Decomposition and other matrix smoothing techniques (Turney and Pantel, 2010). We note that these smoothing techniques are complementary to our method since they could be applied after the context vector combination described in Section 3.2.

The additional use of bilingual (or multilingual) resources for synonym acquisition is also considered in (Van der Plas and Tiedemann, 2006) and (Wu and Zhou, 2003). Their work defines the context of word w in a certain sentence, as the translation of word w, in the corresponding translated sentence. However, their methods require bilingual (or multilingual) parallel corpora. For a word w, they create w's context vector by using all word translations of w, wherein the word translations are determined by the word alignment in the parallel corpus. The weighting of each dimension of the context vector is determined by the number of times word w and its translation are aligned.

The methods described in (Hiroyuki and Morimoto, 2005; Li and Li, 2004) also use comparable corpora and word translations for disambiguating a certain query word. Their methods distinguish word senses by differences in word translations. For example, the senses of plant (factory, vegetation) are distinguished by the translations 工場 (factory) and 植物 (vegetation). Given a text snippet in which the ambiguous word occurs, their methods select the appropriate sense by finding an appropriate translation. In contrast, our method does not use a text snippet to disambiguate the meaning of the query word. Instead, our method uses one or more translations of the query word to find appropriate synonyms. For example, given the query word ”plant” and the translation 工場 (factory) we expect to acquire synonyms like ”manufacture”, ”factory” and so forth.

3 Proposed Method

We assume the user tries to find a synonym for the query term q in language A and provides additional translations of term q in language B. We name these translations as $v_1, ..., v_k$. Furthermore we assume to have a pair of comparable corpora, one in language A and one in language B, and a bilingual dictionary.

We denote $\boldsymbol{q}$ as the context vector of the term q in language A. A context vector $\boldsymbol{q}$ contains in each dimension the degree of association between the term q and another word in language A which occur in the corpus written in language A. Therefore the length of context vector $\boldsymbol{q}$ equals the number of distinct words in the corpus. We will use the notation $\boldsymbol{q}(x)$ to mean the degree of association between the term q and the word x which is calculated based on the co-occurrence of term q and word x in the corpus.

We denote $\boldsymbol{v}_1, ..., \boldsymbol{v}_k$ as the context vectors of the terms $v_1, ..., v_k$ in language B. A context vector $\boldsymbol{v}_i$, $1 \leq i \leq k$, contains in each dimension the degree of association between the term v_i and a word in language B.

3.1 Context Vector Translation

In the first step we estimate the translation probabilities for the words in language B to the words in language A for the words listed in the bilingual dictionary. For that purpose, we build a language model for each language using the comparable corpora, and then estimate the translation probabilities using expectation maximization (EM) algorithm described in (Koehn and Knight, 2000). This way we get the probability that word y in language B has the translation x in language A, which we denote as $p(x|y)$.

We write these translation probabilities into a matrix T which contains in each column the translation probabilities for a word in language B into any word in language A. We use the translation matrix T, in order to translate each vector $\boldsymbol{v}_i$ into a vector which contains the degree of association to words in language A. We denote this new vector as $\boldsymbol{v}_i$, and calculate it as follows:

$$\boldsymbol{v}'_i = T \cdot \boldsymbol{v}_i \tag{1}$$

This way we get the translated context vectors $\boldsymbol{v}'_1, ..., \boldsymbol{v}'_k$.

3.2 Context Vector Combination

In the second step, we combine the context vectors $\boldsymbol{v}'_1, ..., \boldsymbol{v}'_k$ and the context vector $\boldsymbol{q}$. Note that the dimension of a vector $\boldsymbol{v}'_i$ and the vector $\boldsymbol{q}$ is in general different, since $\boldsymbol{v}'_i$ contains only the degree of association to the words listed in the bilingual dictionary.

We could now combine all context vectors additively, similar to monolingual disambiguation like in (Schütze, 1998). However, this would ignore that actually some dimensions are difficult to compare across the two languages. For example, it is difficult to translate the Japanese word かける (hang, put, bring,...) because of its many different meanings depending on the context. Therefore we combine the context vectors to a new context vector $\boldsymbol{q}^*$ as follows: If a word x in language A is in the dictionary, we set

$$\boldsymbol{q}^*(x) := \boldsymbol{q}(x) + \sum_{i=1}^{k} \{(1 - c_x)\boldsymbol{q}(x) + c_x \cdot \boldsymbol{v}'_i(x)\}\,, \tag{2}$$

otherwise we set

$$\boldsymbol{q}^*(x) := (k+1) \cdot \boldsymbol{q}(x)\,. \tag{3}$$

$c_x \in [0, 1]$ is the degree of correspondence between word x and its translations in language B. The intuition of c_x is that, if there is a one-to-one correspondence between x and its translations, then we will set c_x to 1, and therefore consider the context vectors $\boldsymbol{v}'_1$ and $\boldsymbol{q}$ as equally important to describe the degree of association to word x. On the other hand, if there is a many-to-many correspondence, then c_x will be smaller than 1, and we therefore rely more on the context vector of $\boldsymbol{q}$ to describe the degree of association to word x. In case there is no translation available, we can rely only on the context vector of $\boldsymbol{q}$, and therefore set c_x to zero, see Formula (3).

Formally we set c_x as the probability that word x is translated into language B and then back into word x:

$$c_x = p(\bullet|x)^T \cdot p(x|\bullet) \tag{4}$$

where $p(\bullet|x)$ and $p(x|\bullet)$ are column vectors which contain in each dimension the translation probability from word x into the words of language B, and the translation probabilities from words in language B to word x, respectively. These translation probabilities are estimated like Section 3.1.

Finally, note that the vector $\boldsymbol{q}^*$ is not rescaled. Depending on the vector comparison method, it might be necessary to normalize the vector $\boldsymbol{q}^*$. However, we will use in our experiments the cosine similarity to compare two context vectors, so the result does not change if we normalize or rescale $\boldsymbol{q}^*$ by any non-zero factor.

4 Experiments

We extract synonyms from a corpus formed by a collection of complaints concerning automobiles compiled by the Japanese Ministry of Land, Infrastructure, Transport and Tourism (MLIT).[3] Our proposed method additionally consults a comparable corpus which is a collection of complaints concerning automobiles compiled by the USA National Highway Traffic Safety Administration (NHTSA).[4] The Japanese corpus contains 24090 sentences that were POS tagged using MeCab (Kudo et al., 2004). The English corpus contains 47613 sentences, that were POS tagged using Stepp Tagger (Tsuruoka et al., 2005), and use the Lemmatizer (Okazaki et al., 2008) to extract and stem content words (nouns, verbs, adjectives, adverbs).

For creating the context vectors, we use the co-occurrence counts of a word's predecessor and successor from the dependency-parse tree. These co-occurrence counts are then weighted using the

[3] http://www.mlit.go.jp/jidosha/carinf/rcl/defects.html

[4] http://www-odi.nhtsa.dot.gov/downloads/index.cfm

log-odds-ratio (Evert, 2004).[5] For comparing two context vectors we use the cosine similarity. The baseline method is the same as the proposed method except that it does not use Formula (2) and (3) to include information from the translations. As a bilingual dictionary, we use a large-sized Japanese-English dictionary with 1.6 million entries.[6]

In the first experiment we assume that the user wants to acquire all synonyms irrespectively of the difference in senses. Therefore, our gold-standard includes all words which occur in the corpus and which belong to any Japanese WordNet (Bond et al., 2009) synset to which the query term belongs. The gold-standard contains in total 234 Japanese words as query terms.[7] Our proposed method uses as auxiliary translations *all* English translations that correspond the query term and that are listed in our bilingual dictionary. For example, for the query term バルブ (bulb, valve), the proposed method uses the translations ”bulb” and ”valve”.

The results in Table 1 (top) show that in average our method improves finding all synonyms for a query. The improvement can be accounted to the effect that our method enriches the sparse context vector of a Japanese query term.

In our second experiment, we assume that the user is interested only in the synonyms which correspond to a certain sense of the query term. For each query term we include into the gold-standard only the words belonging to one synset, which was randomly chosen. For example, for the ambiguous query term バルブ (bulb, valve) the gold-standard includes only the synset { 弁 (valve)}. That corresponds to a user looking for the synonyms of バルブ (bulb, valve) restricted to the sense of ”valve”. For selecting an appropriate translation, we use the cross-lingually alignment between the synsets of the Japanese and English WordNet (Bond et al., 2009; Fellbaum, 1998). Our proposed method will use as auxiliary translations only the query term's translations *that are listed in the corresponding English synset*. For example, for the query term バルブ (bulb, valve), the proposed method uses only the translation ”valve”.

The results in Table 1 (bottom) show a clear improvement in recall by our proposed method. A pair-wise comparison of our proposed method and the baseline shows a statistically significant improvement over the baseline (p < 0.03).[8] For example. we found that for the query term バルブ (bulb, valve), the baseline ranks 球 (bulb) at rank 3 and 弁 (valve) at rank 4, whereas the proposed method ranked 弁 (valve) at rank 3, and 球 (bulb) at rank 5. This suggests that our method can also help to disambiguate the context vector of an ambiguous query term.

All Senses					
Method	Top 1	Top 5	Top 10	Top 20	Inv. Rank
Baseline	0.10	0.24	0.29	0.37	0.29
Proposed	0.10	0.24	0.33	0.43	0.32

One Sense					
Method	Top 1	Top 5	Top 10	Top 20	Inv. Rank
Baseline	0.10	0.25	0.30	0.37	0.26
Proposed	0.10	0.26	0.35	0.45	0.30

Table 1: Recall at different ranks and inverse rank for gold-standard which considers all senses (top) and only one sense (bottom) for each query term. Recall at rank n is the number of correct synonyms which occur in the list from 1 to n, divided by all correct synonyms for a query. Inverse rank is the sum of the inverse ranks of each correct synonym for a query. All figures are the average over all query terms in the gold-standard.

5 Conclusions

We introduced a new method that combines a query term's context vector with the context vectors of the query term's translations acquired from a comparable corpus. This way our method is able to mitigate problems related to a query term's sparse context vector, and also helps to resolve its ambiguity.

The experiments showed that our method can improve synonym acquisition, when compared to a baseline method which does not use any comparable corpus.

We also demonstrated that our method can help to find the synonyms that are related to only one sense of the query term, by appropriately restricting the query term's translations. This way, our method can also be used to automatically populate resources like WordNet in languages different than English.

[5] In preliminary experiments the log-odds-ratio best among other measures like point-wise mutual information, tf-idf and log-likelihood-ratio.

[6] This bilingual dictionary is not (yet) publicly available.

[7] Each query term as in average 2.3 synonyms which might correspond to different synsets in WordNet. In average, a query term's synonyms belong to 1.2 different synsets.

[8] We use the sign-test (Wilcox, 2009) to test the hypothesis that the proposed method ranks higher than the baseline.

References

L. Bentivogli, I. Dagan, H.T. Dang, D. Giampiccolo, and B. Magnini. 2009. The fifth pascal recognizing textual entailment challenge. *Proceedings of TAC*, 9:14–24.

F. Bond, H. Isahara, S. Fujita, K. Uchimoto, T. Kuribayashi, and K. Kanzaki. 2009. Enhancing the japanese wordnet. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 1–8. Association for Computational Linguistics.

J.R. Curran and M. Moens. 2002. Scaling context space. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 231–238. Association for Computational Linguistics.

S. Evert. 2004. The statistics of word cooccurrences: word pairs and collocations. *Doctoral dissertation, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart*.

C. Fellbaum. 1998. Wordnet: an electronic lexical database. *Cambrige, MIT Press, Language, Speech, and Communication*.

G. Grefenstette. 1994. *Explorations in automatic thesaurus discovery*. Springer.

Kaji Hiroyuki and Yasutsugu Morimoto. 2005. Unsupervised word-sense disambiguation using bilingual comparable corpora. *IEICE transactions on information and systems*, 88(2):289–301.

J. Kazama, S. De Saeger, K. Kuroda, M. Murata, and K. Torisawa. 2010. A bayesian method for robust estimation of distributional similarities. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 247–256. Association for Computational Linguistics.

P. Koehn and K. Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. In *Proceedings of the National Conference on Artificial Intelligence*, pages 711–715. Association for the Advancement of Artificial Intelligence.

T. Kudo, K. Yamamoto, and Y. Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 230–237. Association for Computational Linguistics.

Hang Li and Cong Li. 2004. Word translation disambiguation using bilingual bootstrapping. *Computational Linguistics*, 30(1):1–22.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics.

N. Okazaki, Y. Tsuruoka, S. Ananiadou, and J. Tsujii. 2008. A discriminative candidate generator for string transformations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 447–456. Association for Computational Linguistics.

H. Schütze. 1998. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123.

Y. Tsuruoka, Y. Tateishi, J. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. *Lecture Notes in Computer Science*, 3746:382–392.

P.D. Turney and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

L. Van der Plas and J. Tiedemann. 2006. Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 866–873. Association for Computational Linguistics.

J. Weeds and D. Weir. 2005. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475.

R.R. Wilcox. 2009. *Basic Statistics: Understanding Conventional Methods and Modern Insights*. Oxford University Press.

H. Wu and M. Zhou. 2003. Optimizing synonym extraction using monolingual and bilingual resources. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, pages 72–79. Association for Computational Linguistics.