社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－NL－185 (9)
2008－SLP－ 71 (9)
2008／5／23

単語の類似尺度に基づくシソーラス辞書への用例付与

Nik Adilah Hanin Binti Zahri　　福本 文代
山梨大学大学院医学工学総合研究部
E-mail: {g07mk019 , fukumoto}@yamanashi.ac.jp

本稿では, Linらの類似度尺度により得た単語クラスを用いることで, WordNetシソーラス辞書へReuters'96記事から抽出した用例を付与する手法を提案する. まず, 各単語の出現分布に基づき意味的に類似した単語クラスを作成する. 得られた単語クラスは、単語に関するWordNetシソーラス中の各語義文とその単語を含むReuters'96記事から抽出した各文との類似度を計算する際に利用される. 実験の結果, 10単語からなる名詞単語に対して得られた用例総数1,177文のうち, 正しい語義に割り当てられた用例は964文であり, 81.9%の正解率が得られた.

キーワード: 多義語, 依存関係, 語義, 語義の曖昧性解消, 単語クラスタリング

Example-assignment to WordNet Thesaurus based on Clustering of Similar Words

Nik Adilah Hanin Binti Zahri　　Fukumoto Fumiyo†
Interdisciplinary Graduate School of Medicine and Engineering
University of Yamanashi
E-mail: {g07mk019, fukumoto}@yamanashi.ac.jp

This paper presents a method to retrieve example sentences from Reuters'96 articles to WordNet Thesaurus based on clustering of similar words proposed by (Lin,1998). We first classified words into groups with similar meanings based on distributional pattern of words from corpus. Then, we assigned Reuters'96 sentences which contained each word from corresponding groups to example sentences of each word sense from WordNet Thesaurus by measuring the similarity between both sentences. We evaluated sentences retrieval results against 10 groups of similar words. The evaluation results showed that from 1,177 example sentences retrieved, 964 sentences accommodate the same sense of corresponding words with precision value of 81.9%.

Keywords : polysemous word, dependency triples, word sense, word sense disambiguation, word clustering

1 Introduction

Polysemy is rarely a problem for communication among people. However, polysemy poses a problem in semantic theory and in semantic applications, such as translation or lexicography. Identifying the intended sense of a polysemous word in context is a complicated process as each context alters the sense of the words found in it. Consider the following example of sentence:

(1-1) He published his first *book* in 2006.
(novel, storybook)
(1-2) He jotted something in the *book*.
(notepad, diary)
(1-3) The council had to balance its *books*.
(ledger, account)
(1-4) They run things by the *book* around here.
(law, rule book)

The word "*book*" here represents a few different senses in each sentence according to which word meanings interact when found together in a particular context.

Bootstrapping semantics from text consisting polysemous word is one of the greatest challenges in language technologies especially in machine translation and cross-language information retrieval. The nature of polysemy; forming new senses or larger syntactic units of words when combined with others, often cause the phenomenon of semantic ambiguity. Therefore, the

context in which the word is appeared is important in order to determine the intended sense of word the sentence is referring to.

This paper presents a method for making this initial step. Our goal is to retrieve large number of contexts or example sentences for each word sense by proposing a new method to improve the retrieval of these sentences. Here, we used distributional pattern of words in corpus. Consider the sentence (1-1). The word "*publish*" determines that the sense of "*book*" is referring to "*novel*" or "*storybook*" in the context. Therefore, if the word "*book*" here is replaced by the word "*storybook*" or "*novel*", the meaning of the word sense and the sentence still remain the same. Therefore, we assume that the words with similar meaning share similar context and grammatical relationship in sentences.

In this paper, in order to retrieve large number of sentences for each word sense, we suggest example-assignment to WordNet Thesaurus based on a group of similar words. According to the previous example, similar words that belong to the same sense of word "*book*" here are "*novel*", "*storybook*" or "*tome*". Instead of retrieving sentences that only contains the word "*book*", we also retrieve sentences which contain the words of "*novel*", "*storybook*" and "*tome*" , as shown in the following example:

(2-1) He published his first *book* in 2006.
(2-2) The descriptions in the published *novel* remain superficial.
(2-3) 20 stories were published in our children *storybook*.
(2-4) The company agreed to publish my *tome*.

From this example, we can see that a large number of corresponding contexts or sentences can be retrieved from the same corpus. The next section will describe in detail regarding the similarity measure for similar words clustering. Then, in Section 3, we present a similarity measurement for example-assignment. Next, we review the result of similarity words clustering and example-assignment in Section 4. Section 5 briefly reviews some of previous related works. Finally, Section 6 will discuss our future work in word sense disambiguation (WSD) and summarize our method.

2 Clustering of Similar Words

Our similarity measure for similar word clustering is based on proposal in (Lin,1998) [1]. Precisely, the similarity between two words is measured based on the amount of commonality information in the description shared between the two words. The information of each word is determined according to its grammatical relationship with other words in the text corpus. Thus, we use a broad-coverage parser to extract dependency triple, which consists of two words and a grammatical relationship between them from an input sentence [2]. The example of dependency triples extraction in sentence is shown in Table 1.

The amount of information in the description of a word, w consists of all dependency triples that match the pattern (w,*,*). Here, wild card (*) indicates that the frequency count will include all the triple dependency triples that matches the particular pattern. Let the notation $\|w,r,w'\|$ represents the frequency count of dependency triples (w,r,w'). $\|cook,obj,*\|$, for example, defines the frequency counts of cook-object relationship , and $\|*,*,*\|$ defines the total frequency of dependency triples extracted from the parsed corpus.

The similarity of two words is measured based on commonality of information between those words. Meanwhile, the commonality of information between two words is determined according to the frequency of dependency triples that belong to both words. An occurrence of dependency triple (w, r, w') is composed by the following three co-occurrence events:

A : randomly selected word, w
B: randomly selected dependency type, r
C: randomly selected word,w'

The probability of A,B and C co-occurring is estimated by

$$P_{MLE}(B)P_{MLE}(A|B)P_{MLE}(C|B)$$

Where P_{MLE} is the maximum likelihood estimation of a probability distribution and

$$P_{MLE}(B) = \frac{\| *,r,* \|}{\| *,*,* \|}$$

$$P_{MLE}(A|B) = \frac{\| w,r,* \|}{\| *,r,* \|}$$

$$P_{MLE}(C|B) = \frac{\| *,r,w' \|}{\| *,r,* \|}$$

When the value of $\| w,r,w' \|$ is known, we can obtain $P_{MLE}(A,B,C)$ directly:

$$P_{MLE}(A,B,C) = \frac{\| w,r,w' \|}{\| *,*,* \|}$$

Let $I(w,r,w')$ denotes the amount of information contain in $\|w,r,w'\|$=c and can be computed as follows:

Table 1: Example of dependency triples

Sentence:	He published his first book in 2006
Dependency triples:	(publish, subj, he), (publish, obj, book), (book, gen, his), (book, post, first), (book, mod, in), (in, pcomp-n, 2006)

$$
\begin{aligned}
& I(w, r, w') \\
& = -logP_{MLE}(B)P_{MLE}(A|B)P_{MLE}(C|B) \\
& \quad -(-logP_{MLE}(A, B, C)) \\
& = log\frac{\parallel w, r, w' \parallel \times \parallel *, r, * \parallel}{\parallel w, r, * \parallel \times \parallel *, r, w' \parallel}
\end{aligned}
$$

Let $T(w)$ be the set of pairs (r, w') such that $log\frac{\|w,r,w'\| \times \|*,r,*\|}{\|w,r,*\| \times \|*,r,w'\|}$ is positive. The similarity of two words, $SIM(w_1, w_2)$ is defines as follows;

$$
\frac{\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T(w_2)} I(w_2, r, w)} \quad (1)
$$

We use Eq. (1) to create a WordNet thesauri entries. Table 2 shows some examples of thesaurus entry obtained, which contains the top-5 words that are most similar to each word. Column "Similar Words" shows the top-5 nouns that are computed similar to the noun in "Word" column, ordered in descending order by similarity value. The numbers in the brackets between words are the similarity value computed from Eq. (1).

3 Sentence Retrieval

In order to assign sentences to certain sense of word in the WordNet, we compared Reuters sentences with example sentences acquired from WordNet database. In this section, we will discuss the retrieval of related sentences from Wordnet database and Reuter corpus, and also the similarity measurement used here. First step is the retrieval of sentences for each sense of corresponding word from WordNet database. Next, the same method is applied against Reuters corpus to retrieve sentences contain the same corresponding word and another 5 similar words. These similar words are the top-5 words determined in previous section.

According to examples from (2-1) to (2-4), besides the word "*book*", sentences that contain similar words such as the word "*novel*", "*storybook*" and "*tome*" are also retrieved. Then, word replacement procedure is applied to sentences from Reuter.

Finally, sentences similarity measurement is applied to the sentences obtained by word replacement procedure, and each sentence is assigned to an appropriate sense of word from the WordNet.

3.1 Similar Words Replacement

The similarity measure is computed using sentences retrieved from WordNet database and Reuters. However, before the implementation of similarity measure, we first replaced all the similar words in extracted sentences from Reuters with the word that represents the group. For example, by using the same groups of similar words as mentioned above, the words replacement are implemented as follows:

(3-1) He published his first *book* in 2006.
(3-2) The descriptions in the published ~~*novel*~~ *book* remain superficial.
(3-3) 20 stories were published in our children ~~*storybook*~~ *book*.
(3-4) The company agreed to publish my ~~*tome*~~ *book*.

The purpose of this step is to increase the frequency of one-to-one correspondence words between WordNet and Reuters, which will enhance the value of sentences similarity in similarity measurement.

3.2 Sentence Similarity

We used Brill's tagger to POS-tag both kind of sentences, extract content words and lemmas of the words [4]. Let W_i and R_i be the words of Reuters and WordNet example sentences for the i-th alignment. The similarity between W_i and R_i is defined as follows:

$$
Sent_sim(W_i, R_i) = \frac{co(W_i \times R_i) + 1}{l(W_i) + l(R_i) - 2co(W_i \times R_i) + 2} \quad (2)
$$

where $l(X) = \sum_{x \in X} f(x)$.
$f(x)$ is the frequency of x in the sentence.
$co(W_i \times R_i) = \sum_{(wd,reu) \in W_i \times R_i} min(f(wd), f(reu))$.
$W_i \times R_i = \{(wd, reu) | wd \in W_i, reu \in R_i\}$.

$W_i \times R_i$ is a one to one correspondence between WordNet and Reuters words. We obtained W_i

Table 2: Example of top-5 word entries

Nouns	Similar words
abuse	violation (0.46), crime (0.41), corruption (0.39), fraud (0.39), violence (0.38)
battle	war (0.46), fight (0.46), protest (0.44), dispute (0.44), debate (0.44)
calculation	analysis (0.37), assessment (0.37), valuation (0.35), accounting (0.35), projection (0.34)
debate	discussion (0.49), negotiation (0.48), session (0.47), crisis (0.47), election (0.47)
efficiency	competitiveness (0.44), profitability (0.43), flexibility (0.42), productivity (0.42), quality (0.41)
fair	exhibition (0.38), convention (0.36), seminar (0.35), forum (0.34), tour (0.34)
generator	power plant (0.40), power station (0.37), engine (0.36), refinery (0.36), platform (0.36)
harassment	discrimination (0.40), interference (0.33), violation (0.32), manipulation (0.29), destruction (0.29)
impact	effect (0.54), risk (0.49), pressure (0.48), decline (0.46), change (0.46)
journal	bulletin (0.35), newsletter (0.32), prospectus (0.28), consultancy (0.28), map (0.28)

$\times R_i$, i.e., a one-to-one correspondence between the words in WordNet and Reuters, by looking up the same word in both sentences and measured the similarity between them. According to the similarity result, we only evaluated sentences with similarity that exceed the value of threshold. If the similarity value obtained by Eq. (2) exceeded the threshold value, R_i is regarded having the similar sense of the corresponding word with example sentence of W_i.

4 Evaluation

In this section, we describe our experimental setup and the evaluation results for our system.

4.1 Data

We used Minipar [2], a broad-coverage English parser, to parse 1 year of Reuters'96 data from August 20th, 1996 to August 19th, 1997. This corpus contained 806,791 articles that consist about 9,026,595 sentences. We collected the frequency of dependency triples output by Minipar and used them to compute similar words. Here, we only performed clustering of similar words for nouns. From 806,791 of 1998 Reuters articles, we only extracted *object* and *subject* grammatical relationship of dependency triples. From 289,239 of *object* and *subject* related pairs, there were 30,953 pairs of triple dependency that at least occurred 100 times. Next, we retrieve nouns with frequency 1,000 or higher, and then perform clustering of similar words against them. Here, we obtained 3,167[1] nouns that at least occurred 1,000 times or higher in the parsed corpus. For each noun, we created a thesauri entry which contains the top-5 words that are most similar to it by using the similarity measurement mentioned in Section 2.

[1]Excluding proper nouns

In order to perform sentences similarity measure against Reuters sentences, we used example sentences included in electronic dictionary, WordNet[2] [3]. There are 11,473 of example sentences extracted from WordNet Database Version 3.0. However, in sentence retrieval procedure, only 608 and 180,394 of WordNet and Reuters sentences are used.

4.2 Clustering of Similar Words

From 3,167 group clustered, we randomly selected 25% of groups obtained from similar words clustering to be evaluated manually. We checked if each word belongs to the corresponding clusters and determine the precision value from evaluation result. The sample of evaluation made for a few groups of similar words is shown in Table 3. The bold font word indicates that the word does not belong to its group.

The precision value here is measured by percentage of output clusters that actually correspond to a sense of word. We define the precision of the word clustering is the average precision of all words. As a result, the precision value was 0.330.

The precision value is good considering that clustering result also including the clustering of monosemous words such as numbers or figure. We assume that the value of precision will be higher if the computation excluding the monosemous word.

4.3 Sentences Similarity

For sentences similarity evaluation, we evaluated implemented program by checking the similarity results manually. The evaluation is only performed on sentences with similarity value that exceeded threshold value, $\theta = 0.4$[3]. We randomly

[2]available at http://wordnet.princeton.edu/obtain

[3]The threshold value, θ was empirically determined

Table 3: Samples of word clustering evaluation

Nouns	Similar words
willingness	desire, determination, readiness, intention, commitment
truth	reality, existence, **responsibility**, secret, **violation**
skill	expertise, experience, efficiency, **flexibility**, capability
research	study, work, survey, **marketing**, **development**
obligation	commitment, liability, duty, responsibility, requirement
north	**province, island, town, west, district**
holiday	break, **start**, **auction**, **session**, **entry**
disorder	infection, **outbreak**, illness, **epidemic**, **cancer**
corruption	crime, abuse, violation, violence, disease
allegation	complaint, accusation, **scandal**, claim, case

selected 10 groups of similar words and checked sentences determined by the system to the corresponding group.

At this stage, there was 1,177 of Reuters sentences were manually checked. We considered the similarities calculation for sentences which only contain corresponding word, without sentences that include top-5 similar words, as baseline method for comparison to our method. The amount of sentences obtained from both methods will be compared to determine the precision value for our method. The precision value here is computed by ratio of every measured Reuters sentences whether they accommodate the same sense of corresponding word with WordNet sentences for both methods. The list of words and the amount of sentences involved in the evaluation are listed in the Table 4.

In accordance with data used during evaluation, total sentences that correctly measured for both baseline and our method are 117 and 964 sentences. As can be seen clearly from Table 4, the precision value for baseline is higher, which is P =0.951 against our method, P= 0.819. However, the number of sentences measured by proposed method is 964 sentences, which is definitely 8 times higher than number of sentences retrieved by baseline, 117 sentences. The experimental result showed that our method had significantly improved the sentences retrieval method compared to baseline. Despite of using small number of example sentences from WordNet database, the number of extracted sentences from Reuter was significantly higher when the sentences retrieval was performed according to group based of similar words.

Bootstrapping methods for automatically sense-tag a training corpus has been an interest, as it helps knowledge acquisition bottleneck, i.e., manual sense-tagging of a corpus. The earliest work in this direction are those of Hearst [6] and Yarowsky [7]. Yarowsky's method resolves the problem of knowledge acquisition limitation faced by word-specific sense discriminators disregard the polysemy issues. The identification of rarely occurred word sense in corpus also successfully performed using statically word-specific models. Meanwhile, Gale *et al.* proposed the use of bilingual corpora to avoid hand-tagging of training data. English-French parallel aligned corpus is used to automatically determine sense of each word in target language [8]. This technique is heavily relies on availability of parallel corpora which the main problem to this approach, since the sizes as well as domain of existing bilingual corpora are limited.

Nowadays, language scientists and technologists have a growing tendency to use the Web as a source of language data, due its capability to provides a great amount of linguistic data [7, 9]. Recently, Agirre *et al.* proposed a method to acquire training examples by using two publicly available corpora including Semcor and an additional corpus automatically obtained from the Web [9]. However, they reported that the accuracy using the Web data was decrease, especially when Web examples whose word sense did not appear in publicly corpus. Moreover, the common problem for many Web-based works, the problem of data sparseness has occurred. Our methodology, especially word replacement procedure is the first effort aimed against the problem of data sparseness.

5 Conclusion

Reliable retrieval of example sentences of word sense from text corpus opens up many approaches in the future especially for machine translation and information retrieval systems. This paper presented the initial step to the resolution of lexical semantic ambiguity or known as WSD. In accordance to WSD, our methods were capable to retrieve large number of example sentences

Table 4: Sentences evaluation lists

Word	Sense by WordNet	Number of Sentences Evaluated			
		Baseline		Proposed Method	
		Total	Correct	Total	Correct
penalty	punishment, sentences	7	7	16	10
package	system, programme, deal	37	36	307	234
royalty	commission, bonus	0	0	4	4
rebound	recovery, resurgence	39	38	183	135
examination	test,exam	0	0	16	8
drill	recitation, routine, practice	5	5	29	17
doubt	anxiety, uncertainty, question	8	8	65	65
clearance	permission, approval	11	11	335	301
admission	admittance, entry	1	1	19	19
accord	agreement, settlement	15	11	203	171
Total		123	117	1177	964
Precision		0.951		0.819	

for each sense. The experimental results showed that the number of sentences retrieval by group of similar words was 964 sentences, which was about 8 times higher than baseline method, 117 sentences. The main contribution of this paper is a new method to retrieve sentences for word senses automatically with minimum test data or sentences used for comparison. Our method expands the use of automatic constructed thesauri and helps to develop sentences retrieval for WSD. Moreover, retrieving sentences based on groups of words also capable to generate example sentences for sense which are not provided in dictionaries or thesaurus.

References

[1] D. Lin, "Automatic Retrieval and Clustering of Similar Words", in Proceedings of COLING-ACL '98, pp. 768–774, 1998.

[2] D. Lin, "PRINCIPAR—An Efficient, Broad-coverage, Principle-based Parser", in Proceedings of the 15th International Conference on Computational Linguistics, pp. 42-48, 1994.

[3] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to WordNet: An Online Lexical Database", in Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics, pp. 112-119, 1990.

[4] Eric Brill, "A Simple Rule-based Part-of-Speech Tagger", in Proceedings of 3rd Conference on Applied Natural Language Processing, pp. 152-155, 1992.

[5] M.A.Hearst, "Noun Homograph Disambiguation using Local Context in Large Corpora", in Proceedings of the 7th Annual Conference of the Centre for the New OED and Text Research: Using Corpora, pp. 1-22, 1991.

[6] D. Yarowsky, "Word Sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora", in Proceedings of the 14th International Conference on Computational Linguistics, pp. 454-460, 1992.

[7] R. Mihalcea and I. Moldovan, "An Automatic Method for Generating Sense Tagged Corpora", in Proceedings of the 16th National Conference on Artificial Intelligence, pp. 461-466, 1999.

[8] W.A.Gale and K.W.Church and D.Yarowsky, "Using Bilingual Materials to Develop Word Sense Disambiguation Methods", in Proceedings of the International Conference on theoretical and Methodological Issues in Machine Translation, pp. 101-112, 1992.

[9] E. Agirre and D. Martinez, "Exploring Automatic Word Sense Disambiguation with Decision Lists and the Web", in Proceedings of the 18th International Conference on Computational Linguistics, pp. 11-19, 2000.