

Example-assignment to WordNet Thesaurus based on Distributional Similarity of Words

Fumiyo Fukumoto and Nik Adilah Hanin Binti Zahri†
Interdisciplinary Graduate School of Medicine and Engineering
Univ. of Yamanashi, 4-3-11, Takeda, Kofu, 400-8511, Japan
fukumoto@yamanashi.ac.jp g07mk019@yamanashi.ac.jp†

Abstract. In this paper, we present a method for assigning example sentences to each sense of words in WordNet. The key idea is that the method assigns each sense of a word w collected from not only the sentences containing w , but also sentences containing words with semantically related to w . Because a collection of a *context* of the target word w is a similar syntactic behavior, even collected from a very very large corpora. The evaluation result showed that example-assignment based on groups of similar words significantly improved the retrieval of context for word sense, which helps to determine the exact definition of polysemous words.

1 Introduction

Word Sense Disambiguation (WSD) is one of the important problems in computational linguistics, as it is necessary at one level or another to accomplish most NLP and their applications [20]. One of the major approaches to disambiguate word senses is supervised learning [6], [18], [17]. A typical algorithm constructs a training set from all contexts of a polysemous word in the lexical resources such as machine-readable dictionaries or text corpora, and uses it to learn classifier that maps instances of the polysemous word into the senses [19]. A large number of papers published in this area involve comparisons of different learning approaches trained and tested with commonly used corpora or dictionaries. Unfortunately, not all the semantics are made explicit within lexical resources, even WordNet, the most widespread computational lexicon of English. Moreover, there are not a large amount of example sentences for each sense of words which are used to train classifiers. The production of semantically richer lexical resources can help alleviate the ontology acquisition bottleneck and potentially enable advanced NLP applications. However, in order to reduce the high cost of manual annotation, and to avoid the repetition of this effort for each lexical resource, this task must be supported by wide-coverage automated techniques which do not rely on the specific resource at hand.

In this paper, we present a method for assigning example sentences to each sense of words in WordNet. The key idea is that the method assigns each sense of a word w collected from not only the sentences containing w , but also sentences containing words with semantically related to w . Because a collection of a *context* of the target word w is a similar syntactic behavior, even collected from a very very large corpora [2]. Here, a *context* of the target word w is any sentence that contains w in the corpus. Consider the word “book” in WordNet. It has at least five senses including *account* sense. We would use, in addition to the contexts of “book”, all the contexts of “account” in

the Reuters’96. The word “book” has a sense “account” when it co-occurred with the word “balance”. However, co-occurrences between “book” and “balance” are not observed even in a very large corpus, while those between “account” and “balance” are high. For instance, in a corpus collected from one month Reuters, the number of sentences that contains “book” and “balance” were only 6, while those of “account” and “balance” were 65. Therefore, if we can find that “book” and “account” are semantically related, and collect contexts containing “account” and “balance”, it would be possible to improve disambiguation accuracy using only seed annotated sentences *i.e.*, example sentences in WordNet together with a large unlabeled corpus, without requiring any additional hand labeling.

2 Overview of the System

The method consists of two steps: collecting semantically similar words and sentence retrieval. Figure 1 illustrates an overview of the system.

2.1 Collection of Semantically Similar Words

The first step to assign example sentences to WordNet thesaurus is to collect similar words from a corpus. A typical measure of similarity between words is based on their distributional similarity [9], [3]. Similarity measures based on distributional hypothesis compare a pair of weighted feature vectors that characterize two words. Features typically correspond to other words that co-occur with the characterized word in the same context. It is then assumed that different words that occur within similar contexts are semantically similar. Lin proposed a word similarity measure based on the distributional pattern of words which allows to construct a thesaurus using a parsed corpus [12]. We used it to calculate word similarities. More precisely, the similarity between two words is measured by the ratio between the amount of information needed to state the commonality of two words and the information needed to fully describe what the two words are [11]. It is based on their grammatical relationship with other words in the text corpus. We used Lin’s syntactic parser to extract dependency triples [10]. The example of dependency triples is shown in Table 1.

The amount of information of a word w consists of all dependency triples that matches the pattern $(w, *, *)$, where wild card (*) indicates frequencies including all the 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

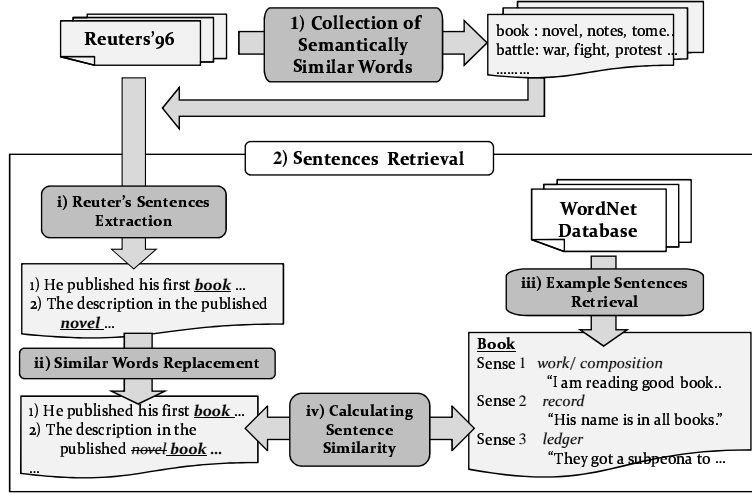


Figure 1. Overview of the System

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)

$\|*,*,*\|$ defines the total frequency of dependency triples extracted from the parsed corpus.

The similarity of two words is measured based on the frequency of dependency triples. 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

$$\begin{aligned} P_{MLE}(B) &= \frac{\|*,r,*\|}{\|*,*,*\|}, \\ P_{MLE}(A|B) &= \frac{\|w,r,*\|}{\|*,r,*\|}, \\ P_{MLE}(C|B) &= \frac{\|*,r,w'\|}{\|*,r,*\|}. \end{aligned} \quad (1)$$

P_{MLE} is the maximum likelihood estimation of a probability distribution. When the value of $\|w,r,w'\|$ is known, we can obtain $P_{MLE}(A, B, C)$ directly:

$$P_{MLE}(A, B, C) = \frac{\|w,r,w'\|}{\|*,*,*\|} \quad (2)$$

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

$$\begin{aligned} I(w, r, w') &= -\log P_{MLE}(B)P_{MLE}(A|B)P_{MLE}(C|B) \\ &\quad -(-\log P_{MLE}(A, B, C)), \\ &= \log \frac{\|w, r, w'\| \times \|*, r, *\|}{\|w, r, *\| \times \|*, r, w'\|}. \end{aligned} \quad (3)$$

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 defined as:

$$\frac{\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T(w_2)} I(w_2, r, w)}. \quad (4)$$

A WordNet thesauri entry is created by using Eq. (4). For each noun word (we call it *seed* word), we extracted the top-5 words with the highest similarity value as a group of similar words.

2.2 Sentence Retrieval

The second step is to retrieve example sentences from Reuters, and assign each sentence to each sense of words in the WordNet. This step consists of two sub-steps: similar words replacement and calculating sentence similarity. In the similar words replacement, we replaced all the similar words in the extracted sentences from Reuters with seed word. The purpose of this sub-step is to increase the frequency of one-to-one correspondence words between WordNet and Reuters, which will extend the value of sentences similarity in the next sub-step. Next, we calculate two sentences from Reuters and WordNet examples by using formula (5).

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

where

$$|X| = \sum_{x \in X} f(x),$$

$$co(W_i \times R_j) = \sum_{(wn, reu) \in W_i \times R_j} \min(f(wn), f(reu)),$$

$$W_i \times R_j = \{(wn, reu) | wn \in W_i, reu \in R_j\}. \quad (5)$$

$f(x)$ denotes the frequency of x in the sentence X . In Eq. (5), W_i and R_j refer to a set of words of the i -th Reuters sentence and j -th WordNet example sentence, respectively. (wn, reu) refers to one-to-one correspondence between the words wn and reu by looking up the same word in both sentences or words within the same group. If the similarity value exceeded a certain threshold value, R_i is regarded having the similar sense of the corresponding word with an example sentence of W_i .

3 Evaluation

3.1 Data

We used Minipar [10], a broad-coverage English parser, to parse 1 year of Reuters'96 data from August 20th, 1996 to August 19th, 1997. These corpus contained 806,791 articles consisting of 9,026,595 sentences. We collected the frequency of dependency triples by Minipar and used them to collect similar words. Here, we only performed clustering of similar noun words. From 806,791 Reuters articles, we extracted *object* and *subject* grammatical relationship of dependency triples. From 289,239 of *object* and *subject* related pairs, we obtained 30,953 pairs of triple dependency that occurred at least 100 times. Next, we retrieved 3,167 nouns with frequency 1,000 or higher, and then performed clustering of similar words against them. For each noun, we created a thesauri entry which contains the top-5 words that are most similar to the *seed* word by using the similarity measure mentioned in Section 2.1.

In order to perform sentences similarity measure against Reuters sentences, we used example sentences in WordNet² [14]. There are 11,473 example sentences extracted from WordNet Database Version 3.0. However, in sentence retrieval procedure, we used only 608 WordNet and 180,394 of Reuters sentences.

3.2 Collection of Similar Words

From 3,167 group clustered, we randomly selected 25 percent of groups: 792 groups obtained from similar words clustering to be evaluated them manually. We checked if each word belongs to the corresponding groups. The sample of evaluation for the groups of similar words is shown in Table 2. The bold font word indicates that the word does not belong to its group.

We can see from Table 2 that some groups such as “willingness” and “obligation” were perfectly clustered, while “north” group consists of different sense of words. Table 3 shows the distributional of 792 groups. “# number of words correctly clustered” refers to the number of words that actually belong to the corresponding group determined by the system, and “# number of groups” denotes the amount of group evaluated with corresponding number of correctly determined similar words. Table 3 shows that the total number of similar words correctly identified by the system was 1,315, which resulted precision value, $P=0.332$.

3.3 Sentence Retrieval

We evaluated 1,629 sentences with similarity that exceeded the threshold value, $\theta = 0.3$. The list of words and the amount of sentences evaluated are listed in the Table 4. “baseline” in Table 4 shows

² available at <http://wordnet.princeton.edu/obtain>

Table 2. Samples of word clustering evaluation

Nouns	Similar words
access	use, control, link, privatization, standard
acceptance	recognition, integration, creation, efficiency, participation
accord	pact, treaty, agreement, legislation, measure
council	commission, committee, parliament, cabinet, agency
corruption	crime, abuse, violation, violence, disease
criticism	threat, speculation , complaint, allegation , comment
discussion	negotiable, debate, privatization , investigation, study
disorder	infection, outbreak , illness, epidemic, cancer
fluctuation	appreciation, downturn, swing, slump, instability
holiday	break, start, auction, session, entry
match	game, championship, round, race, final
north	province, island, town, west, district
regulation	rule, legislation, law, standard, pact
view	expectation, statement, term, report, idea
willingness	desire, determination, readiness, intention, commitment

Table 3. Distributional of evaluation results

# of Words Correctly Clustered in a Group	# of Groups	Total # of Correctly Clustered Words
0	210	0
1	192	192
2	182	364
3	106	318
4	69	276
5	33	165
Total	792	1,315

the results obtained by sentence similarity measure which does not involve in word replacement procedure.

As can be seen clearly from Table 4, the precision value for baseline was higher, which was $P = 0.828$ against our method, $P = 0.770$. However, the number of sentences measured by our method was 1,254 sentences, which were definitely 8 times higher than the number of sentences retrieved by the baseline, 159 sentences. Moreover, we found that our method retrieved sentences for the same sense of any top-5 similar words which did not exist in WordNet database. Table 5 shows an example taken from the sentences for the word “accord” with *settlement* and *agreement* sense. We can see from Table 5 that our method retrieves 61 sentences that contain “pact”, 54 sentences containing “treaty”, and 21 sentences that contain “legislation” with the same sense of the word “accord”. Some of the example sentences are:

- Reu960826-18900:* There was an iron **pact** at work here between certain political party.
- Reu961018-22517:* The **pact** must be qualitative, and not quantitative.
- Reu970507-18324:* Detail of the **pact** is not immediately available.
- Reu961007-2549:* The **treaty** was nonetheless being observed by all party.
- Reu970512-25674:* “the **treaty** is flexible enough,” he said.
- ...

We also found that some sentences are too general and did not show any specific feature to differentiate multiple sense of corresponding word. Consider the following sentences taken from the re-

Table 4. Evaluation result of sentence similarity

Word	Sense by WordNet	The number of sentences evaluated			
		Baseline		Proposed method	
		Correct	Total	Correct	Total
access	approach	0	0	2	2
	act of entering	0	0	3	4
accord	agreement, settlement	31	33	688	834
admission	admittance, entry	19	25	250	351
disaster	catastrophe, ruin, mess, misfortune	7	7	24	29
	tragedy, calamity (events cause by nature)	1	1	3	3
discussion	give-and-take word	10	10	18	21
	treatment, discourse	0	0	2	4
drill	exercise, practice	19	43	99	169
fluctuation	a wave motion	7	7	18	22
penalty	punishment, sentences	61	62	108	151
regulation	rule (principle or condition)	3	3	36	36
royalty	bonus, commission	1	1	3	3
Total		159	192	1,254	1,629
Precision		0.828		0.770	

Table 5. Example of sentences generation

	Word	Top-5 similar words					Total
	accord	pact	treaty	agreement	legislation	measure	
Example sentences (WordNet)	2	0	0	6	0	8	16
# of sentences	31	61	54	463	21	58	688

sults.

- Reu970203-9425:* This would only create an enormous **fluctuation**.
- Reu970312-22347:* This be a short-term **fluctuation**.
- Reu970312-22347:* It was just a normal **fluctuation**.

According to WordNet Thesaurus, the word “*fluctuation*” is either a physical wave motion or unsteady and changing condition. These three sentences have both senses, and consequently some of the sentences retrieved did not have the target sense in the sentences.

4 Previous Work

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 [8], Schütze [16] and Yarowsky [18]. 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 by using statically word-specific models. 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 [6]. One problem with this approach is that the techniques heavily rely on availability of parallel corpora, while the sizes as well as the domain of existing bilingual corpora are limited. Dagan proposed a similar method, but instead of a parallel corpus use two monolingual corpora and a bilingual dictionary [5]. This solves the problems of availability of parallel corpora, since monolingual

corpora are much easier to obtain than parallel corpora.

More recently, language scientists and technologists are increasingly turning to the Web as a source of language data, because it provides a great big body of linguistic data [18], [1], [13]. Agirre *et al.* proposed a method to acquire training examples by using two publicly available corpora including Semcor and an additional corpus automatically acquired from the Web [1]. 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 problem of data sparseness, which is especially severe for work in WSD occurred. There are at least three techniques to solve the problem of data sparseness: smoothing, class-based, and similarity-based methods. Smoothing method is used to get around the problem of infrequently occurring events, and in particular to ensure that non-observed events are not assumed to have a probability of zero. The best-known smoothing method is Turing-Good [7]. Class-based model is a method to obtain the best estimates by combining observations of classes of words considered to belong to a common category. Resnik used the taxonomy of WordNet and Yarowsky used the categories of *Roget’sThesaurus* to define classes [15], [18]. Similarity-based method is to estimate similar words of the target word by using some similarity metric between patterns of co-occurrence [4]. Our methodology, especially word replacement procedure uses similarity-based method which makes it possible to assign sentences not containing the target word w to each sense of w .

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 the context of WSD, our methods retrieved large number of example sentences for each sense. The experimental results showed that the number of sentences retrieval by group of similar words was 1,254 sentences, which was 8 times higher than baseline method, 159 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, we found that our method retrieved sentences for the same sense of any top-5 similar words which did not exist in WordNet database. Future work will include (i) applying the method to other thesaurus such as Roget's thesaurus and LDOCE, and (ii) applying the method to WSD task.

REFERENCES

- [1] 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).
- [2] M. Banko and E. Brill, 'Scaling to very very large corpora for natural language disambiguation', in *Proc. of the 39th Annual Meeting and 10th conference of the European Chapter*, pp. 26–33, (2001).
- [3] I. Dagan, L. Lee, and F. C. N. Pereira, 'Similarity-based models of word cooccurrence probabilities', *Machine Learning*, **34**(1-3), 43–69, (1999).
- [4] I. Dagan, S. Marcus, and S. Markovitch, 'Contextual Word Similarity and Estimation from Sparse Data', in *Proc. of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 164–171, (1993).
- [5] I. Dagan, F. Peireira, and L. Lee, 'Similarity-based Estimation of Word Cooccurrence Probabilities', in *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 272–278, (1994).
- [6] W. A. Gale, K. W. Church, and D. Yarowsky, 'Using bilingual materials to develop word sense disambiguation method', in *Proc. of the International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 101–112, (1992).
- [7] I. J. Good, 'The Population Frequencies of Species and the Distribution of Population Parameters', in *Biometrika*, pp. 237–264, (1953).
- [8] 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).
- [9] D. Hindle, 'Noun classification from predicate-argument structures', in *Proc. of 28th Annual Meeting of the Association for Computational Linguistics*, pp. 268–275, (1990).
- [10] D. Lin, 'Principar—an efficient broad-coverage principle-based parser', in *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 42–48, (1994).
- [11] D. Lin, 'Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity', in *Proc. of 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 64–71, (1997).
- [12] D. Lin, 'Automatic Retrieval and Clustering of Similar Words', in *Proc. of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pp. 768–773, (1998).
- [13] R. Mihalcea, 'Using wikipedia for automatic word sense disambiguation', in *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL2007)*, pp. 196–203, (2007).
- [14] 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).
- [15] P. Resnik, 'WordNet and Distributional Analysis: A Class-based Approach to Statistical Discovery', in *Proceedings of the AAAI Workshop on Statistically-based Natural Language Processing Techniques*, pp. 48–56, (1992).
- [16] H. Schütze, 'Dimensions of Meaning', in *Proc. of Supercomputing'92*, pp. 787–796, (1992).
- [17] M. Stevenson and Y. Wilks, 'The interaction of knowledge sources in word sense disambiguation', *Computational Linguistics*, **27**(3), 321–350, (2001).
- [18] D. Yarowsky, 'Word sense disambiguation using statistical models of roget's categories trained on +arge corpora', in *Proc. of the 14th International Conference on Computational Linguistics*, pp. 454–460, (1992).
- [19] W. Yorick, D. Fass, C. M. Gao, J. E. McDonald, T. Plate, and B.M. Slator, 'Providing machine tractable dictionary tools', *Machine Translation*, **5**(2), 99–154, (1990).
- [20] W. Yorick and M. Stevenson, 'The grammar of sense: Is word sense tagging much more than part-of-speech tagging?', in *Technical Report CS-96-05(University of Sheffield)*, (1996).