Constructing Word-Sense Association Networks from Bilingual Dictionary and Comparable Corpora

Hiroyuki Kaji and Osamu Imaichi

Central Research Laboratory, Hitachi, Ltd.
1-280 Higashi-Koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan
{kaji, imaichi}@crl.hitachi.co.jp

Abstract

A novel thesaurus named a "word-sense association network" is proposed for the first time. It consists of nodes representing word senses, each of which is defined as a set consisting of a word and its translation equivalents, and edges connecting topically associated word senses. This word-sense association network is produced from a bilingual dictionary and comparable corpora by means of a newly developed fully automatic method. The feasibility and effectiveness of the method were demonstrated experimentally by using the EDR English-Japanese dictionary together with *Wall Street Journal* and *Nihon Keizai Shimbun* corpora. The word-sense association networks were applied to word-sense disambiguation as well as to a query interface for information retrieval.

1 Introduction

Word associations are useful for solving some of the important problems in natural language processing. For example, they provide effective clues for parsing ambiguous structures in sentences. However, the usefulness of word associations is in their potential for solving other problems like word-sense disambiguation, since word senses relevant to each word association are implicit.

We propose a novel thesaurus named a word-sense association network, which consists of word-sense associations, not word associations. Then, we develop a method for producing a word-sense association network automatically from a bilingual dictionary and a pair of comparable corpora. Here, we focus on using weakly comparable corpora, i.e., texts of the same genre or domain written in different languages, taking into account the limited availability of large parallel or tightly comparable corpora.

So-called association thesauri, which are automatically produced from corpora, have been used particularly in information retrieval (Jing and Croft 1994; Schuetze and Pedersen 1994; Mandala, et al. 1999; Kaji, et al. 2000). The difference between association thesauri and word-sense association networks is that the former are collections of word associations, while the latter are collections of word-sense associations. Taxonomy-type thesauri, including WordNet (Miller 1990), which are complementary to the word-sense association networks, have been constructed manually. MindNet (Richardson, et al. 1998) should be mentioned specially since it was produced automatically and includes labeled word-sense associations. While the method developed to produce MindNet was applicable to machine-readable dictionaries such as LDOCE, we aim at developing a method for acquiring word-sense associations from general texts so that wide-coverage networks can be constructed.

2 What is Word-Sense Association Network?

A word-sense association network is a network-type thesaurus, where a node represents a word sense and an edge connects a pair of topically associated word senses. Senses of a word are defined as synonym sets consisting of the word and some of its translation equivalents in another language. For example, the sense of the English word "tank" as a military vehicle is defined as {tank, 戦車<*SENSHA*>}, while its sense as a container is defined as {tank, タンク<*TANKU*>, 水槽<*SUISOU*>}. The sense {tank, 戦車<*SENSHA*>} is connected to senses such as {soldier, 兵士<*HEISHI*>} and {troop, 隊<*TAI*>, 軍隊<*GUNTAI*>}, which are topically associated with it.

A crucial issue in specifying a sense-based thesaurus is how to define senses of a word. Assuming that polysemy is not parallel between languages, we adopted the above-described method. This assumption is true for most polysemous words between languages with different origins like English and Japanese (Resnik and Yarowsky 2000). For example, "戦車<*SENSHA*>" is a translation equivalent for "tank" as a military vehicle, but not for "tank" as a container. Defining word senses with translation equivalents has two advantages. First, (bilingual) people can easily identify the exact senses from definitions. Second, definitions can be generated automatically as described in the following section.

3 How is a Word-Sense Association Network Produced?

3.1 Overview of proposed method

The basic idea for the proposed method is to align first-language pairs of associated words with second-language pairs of associated words by consulting a bilingual dictionary, which results in pairs of associated word senses. It is common to methods of word-sense disambiguation using comparable corpora (e.g., Dagan and Itai 1994). This naive idea, however, encounters the following problems. First, aligning pairs of associated words suffers from ambiguity of the alignment suggested by the bilingual dictionary as well as failure to align due to incompleteness of the bilingual dictionary and disparity in the topical coverage between the corpora of the two languages in question. Second, since the correspondence between senses and translation equivalents is one-to-many, clustering translation equivalents needs to be done simultaneously to generate definitions of senses. Third, a method must be computationally efficient so that it can deal with the whole vocabulary of the two languages.

To overcome these problems, the basic idea was elaborated into the method shown in Figure 1, which consists of the following steps.

(1) Extract word associations

Word associations, i.e., pairs of associated words, are extracted from a corpus in each language. That is, mutual infor-

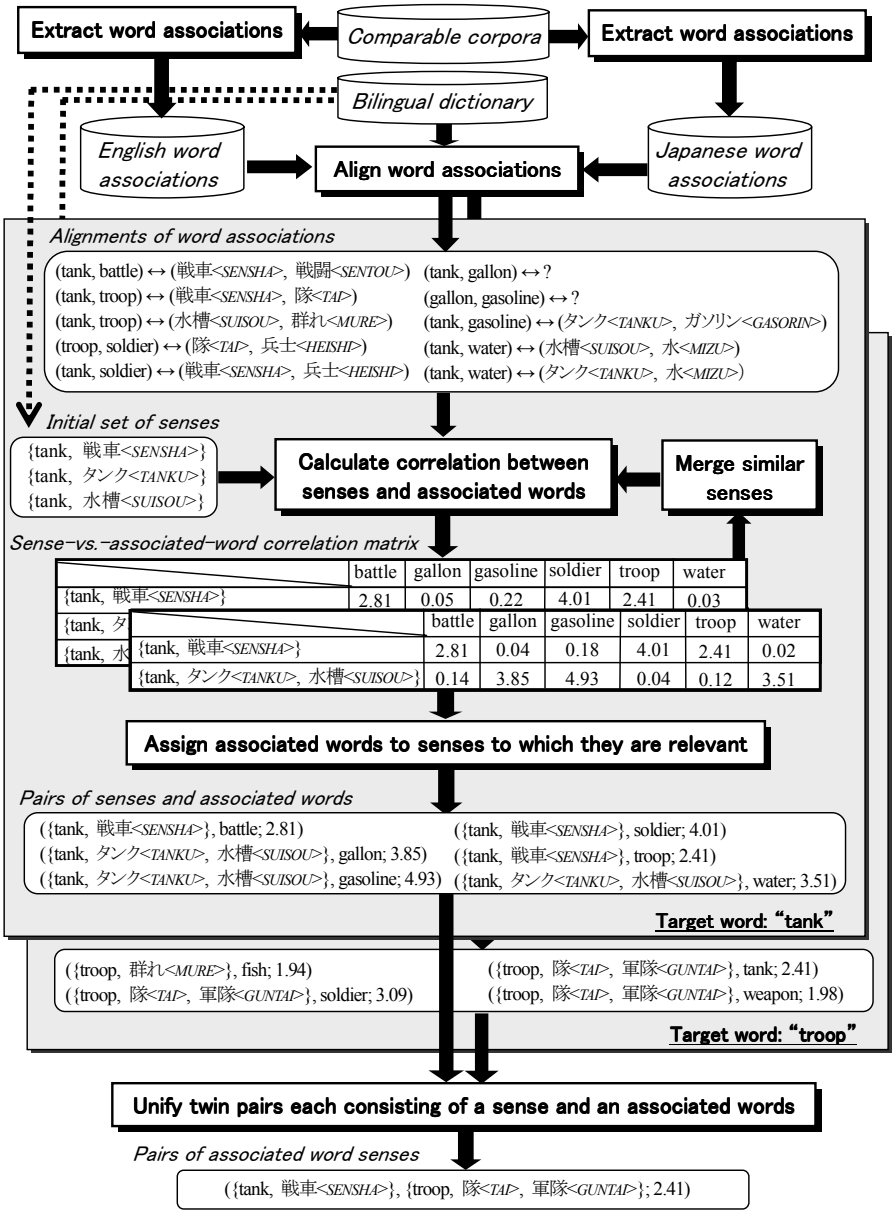


Figure 1: Flow diagram of proposed method for producing a word-sense association network

mation is calculated for every pair of words according to their co-occurrence frequency and respective occurrence frequencies, and pairs of words with mutual information larger than a threshold are collected. For example, a collection of word associations including (tank, soldier) and (tank, gasoline) is extracted from an English corpus, and another collection of word associations including (戦車<*SENSHA*>, 兵士<*HEISHI*>) and (タンク<*TANKU*>, ガソリン<*GASORIN*>) is extracted from a Japanese corpus comparable to the English corpus.

(2) Align word associations

Word associations are aligned translingually by consulting a bilingual dictionary. For example, (tank, soldier) and (tank, gasoline) are aligned with (戦車<*SENSHA*>, 兵士<*HEISHI*>) and (タンク<*TANKU*>, ガソリン<*GASORIN*>), respectively. Note that a word association is aligned with all counterparts suggested by the bilingual dictionary. For example, (tank, troop) is aligned both with (戦車<*SENSHA*>, 隊<*TAI*>) and with (水槽<*SUISOU*>, 群れ<*MURE*>).

(3) Extract sense and associated-word pairs

For each target word, pairs consisting of one of its senses and one of its associated words are extracted according to the aligned word associations. This step consists of the following three substeps (3a-3c). Note that the first two subsetps are repeated alternately and result in a set of senses of the target word as well as a sense-vs.-associated-word correlation matrix. The initial set of senses consists of ones defined with respective translation equivalents; thus, a sense may initially be defined in duplicate.

(3a) Calculate correlation between senses and associated words

A correlation between a sense and an associated word is defined as the mutual information between the target word and the associated word multiplied by a plausibility factor, which reflects the degree of correlation of the sense with the accompanying associated words (i.e., other associated words that are associated with the associated word concerned). A sense-vs.-associated-word correlation matrix is calculated iteratively according to this recursive definition of correlation.

(3b) Merge similar senses

Since senses are characterized by respective rows in the sense-vs.-associated-word correlation matrix, senses with high-similarity row vectors are merged into one.

(3c) Assign associated words to senses to which they are relevant

Each associated word is assigned to the sense having the highest correlation with the associated word, resulting in a set of pairs of senses and associated words for the target word. In Figure 1, for example, pairs such as ({tank, 戦車<*SENSHA*>}, soldier) and ({tank, タンク<*TANKU*>, 水槽<*SUISOU*>}, gasoline) are extracted for the target word "tank."

(4) Unify twin pairs each consisting of a sense and an associated word

From all pairs of senses and associated words extracted for all target words, twin pairs, in which the target word and the associated word are exchanged, are unified as a pair of associated word senses. For example, ({tank, 戦車<*SENSHA*>}, troop) and ({troop, 隊<*TAI*>, 軍隊<*GUNTAI*>}, tank), which are extracted for the target words "tank" and "troop" respectively, are unified as ({tank, 戦車<*SENSHA*>}, {troop, 隊<*TAI*>, 軍隊<*GUNTAI*>}).

3.2 Main features of proposed method

3.2.1 Iterative calculation of correlation between senses and associated words (Kaji and Morimoto 2002)

The recursive definition of correlation between senses and associated words was devised to overcome the ambiguity in alignment of word associations as well as to recover word associations that fail to be aligned with their counterparts. Accompanying associated words play an essential role in the iterative process of calculating the correlations. In Figure 1, (tank, troop), which is aligned both with (戦車<*SENSHA*>, 隊<*TAI*>) and with (水槽<*SUISOU*>, 群れ<*MURE*>), results in ({tank, 戦車<*SENSHA*>}, troop), because the sense {tank, 戦車<*SENSHA*>} has higher correlation with the accompanying associated word "soldier" than the other sense {tank, タンク<*TANKU*>, 水槽<*SUISOU*>}. Additionally, (gallon, gasoline), which is not aligned with any word association in Japanese, results in ({tank, タンク<*TANKU*>, 水槽<*SUISOU*>}, gallon), because {tank, タンク<*TANKU*>, 水槽<*SUISOU*>} has higher correlation with the accompanying associated word "gasoline" than {tank, 戦車<*SENSHA*>}.

3.2.2 Word-sense clustering based on correlation between senses and associated words (Kaji 2003)

Clustering senses corresponds to clustering translation equivalents, since each sense is defined using one or more translation equivalents. We call the proposed method "translingual distributional word clustering," since it clusters second-language translation equivalents characterized by first-language associated words. Ordinary distributional word clustering (e.g., Pereira et al. 1993) is obviously an alternative to the proposed method; translation equivalents are clustered according to their associated words in their own language. Translingual distributional clustering has a number of advantages over the alternative. First, it allows corpus-irrelevant translation equivalents to be filtered out from clustering results. Second, it alleviates the sparseness problem of word-association data owing to the smoothing effect of the iterative calculation of correlation between senses and associated words.

3.2.3 Separate handling of respective target words

The essential part of the proposed method, i.e., extracting pairs consisting of a sense and an associated word, is executed for each target word. It calculates a rather small correlation matrix of at most a few dozen senses vs. a few hundred associated words. This makes the proposed method computationally efficient. Note that Tanaka and Iwasaki's (1996) translation-probability matrix optimization method, which is based on a similar idea as ours, is hampered by the huge amount of computation required to optimize a matrix made up of first-language vocabulary vs. second-language vocabulary. It should be noted that the proposed method allows a word-sense association network to be constructed accumulatively, namely, enlarging target words from frequently-occurring words to less-frequently-occurring words.

4 Experiment

An experiment was done by using the EDR (Japan Electronic Dictionary Research Institute) English-Japanese dictionary, an English corpus consisting of *Wall Street Journal* articles (July 1994 to December 1995; 189MB), and a Japanese corpus consisting of *Nihon Keizai Shimbun* articles (December 1993 to November 1994; 275MB). It proved the computational feasibility and effectiveness of the proposed method as described in the following, although the quality of the resulting word-sense association network remains to be evaluated in detail.

We focused on nouns, including compound nouns. A window of 13 words excluding function words was used to count co-occurrence frequencies, and pairs of nouns with mutual information larger than zero were extracted as word associations. Then, 5,000 most-frequently-occurring words were taken as target words from 269,000 English nouns included in the EDR English-Japanese dictionary, and pairs consisting of a sense of each target word and its associated word sense (which was not limited to the senses of the target words) were acquired.

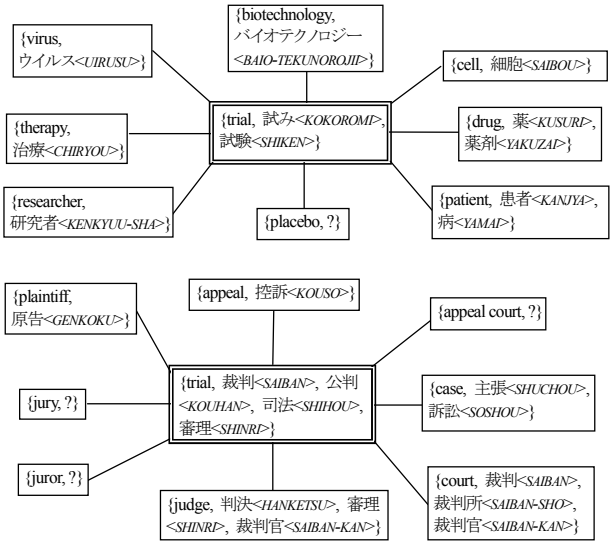
The resulting word-sense association network contained 8,217 word senses derived from the 5,000 target words and additional word senses derived from non target words. The total number of word-sense associations was 715,475, of which those consisting of word senses derived from the target words were counted in duplicate.

Examples of acquired word senses and word-sense associations are shown in Figure 2. For a target word "trial," two senses were acquired: one is a process of testing to find out whether something works effectively, and the other is a legal process in which a court of law examines a case. For each of these senses, eight most representative associated word senses are shown in the figure. The network shows that the sense of "appeal" as a formal request to a court, the sense of "court" as a law court, and others are associated with the legal sense of "trial." The representativeness of associated word senses was evaluated according to how many closely associated word senses accompany each associated word sense.

In addition, the processing time required to produce the network was measured. It took 21.0 hours on a Windows PC (CPU clock: 2.40 GHz; memory: 512 MB) to process the 5,000 target words. The processing times for the first 100 and last 100 target words were 5.7 hours and 4.0 minutes, respectively; additional target words would thus be processed more efficiently. This is because pairs consisting of a sense and an associated word once extracted are saved, and they are retrieved when their counterparts are extracted.

5 Applications

Among various applications of word-sense association



[Note: A question mark is shown in the part of translation equivalents to define a sense, when the original bilingual dictionary gives no translation equivalents.]

Figure 2: Part of a produced word-sense association network

networks, word-sense disambiguation (WSD) and a query interface for information retrieval (IR) are discussed in the following. Other possible applications include synonym identification and improved query expansion.

5.1 Word-sense disambiguation

WSD using a word-sense association network is straightforward. That is, for each occurrence of a polysemous word in a text, the score of each sense is calculated according to its associated words occurring in the neighborhood, and the sense with the highest score is selected. An evaluation experiment, where definitions of senses were given manually, has shown a promising performance; namely, the *F*-measure was 74.6% averaged over 60 typical English polysemous words, compared to a 62.8% baseline by the most-frequent-sense selection method (Kaji and Morimoto 2002).

5.2 Word sense-based query interface

Navigating in a word-sense association network allows IR-system users to articulate their information needs and specify unambiguous queries. Not only bilingual people but also monolingual people can identify the sense each node in the network represents, thanks to the nodes it is connected to. In Fig. 2, for example, even English-monolingual people can understand that {trial, 裁判<*SAIBAN*>, …} means a legal process, because it is connected to {appeal, 控訴<*KOUSO*>}, {appeal court, ?}, {case, 主張<*SHUCHOU*>, …}, and others. The word-sense association network thus provides a query interface especially useful for cross-language IR. We have developed a prototype word-sense-association-network navigator, which functions as a front-end processor for Web-search engines.

6 Conclusion

A sense-based thesaurus named a word-sense association network was proposed, and a fully automatic method to produce it from a bilingual dictionary and a pair of weakly comparable corpora was developed. We expect that word-sense association networks, which facilitate processing of lexical polysemy and synonymy, will play essential roles in many applications of natural language processing.

We are planning to extend word-sense association networks in several ways. One of them will be to combine two networks produced by exchanging the first and second languages. Another interesting issue is to integrate WordNet and a word-sense association network, which are complementary to one another. Note that definitions of senses with synonymous translation equivalents are compatible with WordNet synsets.

Acknowledgments

We are grateful for the valuable discussions with and comments given by Prof. Toyoaki Nishida and Prof. Jun'ichi Tsujii of the University of Tokyo. This research was sponsored by the New Energy and Industrial Technology Development Organization of Japan (NEDO).

References

Dagan, Ido and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4): 563-596.

Jing, Yufeng and W. Bruce Croft. 1994. An association thesaurus for information retrieval. In *Proceedings of a Conference on Intelligent Text and Image Handling "RIAO'94"*, pages 146-160.

Kaji, Hiroyuki, Yasutsugu Morimoto, Toshiko Aizono, and Noriyuki Yamasaki. 2000. Corpus-dependent association thesaurus for information retrieval, In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 404-410.

Kaji, Hiroyuki and Yasutsugu Morimoto. 2002. Unsupervised word sense disambiguation using bilingual comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 411-417.

Kaji, Hiroyuki. 2003. Word sense acquisition from bilingual comparable corpora. In *Proceedings of HLT-NAACL 2003, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 111-118.

Mandala, Rila, Takenobu Tokunaga, and Hozumi Tanaka. 1999. Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development of Information Retrieval*, pages 191-197.

Miller, George A. 1990. WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4): 235-312.

Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183-190.

Resnik, Philip and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2): 113-133.

Richardson, Stephen D., William B. Dolan, and Lucy Vanderwende. 1998. MindNet: acquiring and structuring semantic information from text. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 1098-1102.

Schuetze, Hinrich and Jan O. Pedersen. 1994. A cooccurrence-based thesaurus and two applications to information retrieval. In *Proceedings of a Conference on Intelligent Text and Image Handling "RIAO'94"*, pages 266-274.

Tanaka, Kumiko and Hideya Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora, In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 580-585.