Automatic Construction of

Word Sense Association Networks

Hiroyuki Kaji

September 2003

Contents

Chapter 1 Introduction **1**

1.1 Motivation 1

1.2 Purpose of study 2

1.3 Survey of previous work in related areas 4

1.3.1 Semantic lexicons 4

1.3.2 Translation equivalent extraction 5

1.3.3 Corpus-based word sense disambiguation 6

1.3.4 Distributional word clustering 8

1.4 Thesis overview 9

Chapter 2 From Word Associations to Word Sense Associations: An Approach Using Bilingual Corpora **10**

2.1 What is a word sense association network? 10

2.1.1 Word association: its usefulness and limitations 10

2.1.2 Definition of word senses 11

2.1.3 Word sense association network 13

2.2 How word sense association networks are produced? 16

2.2.1 Basic idea 16

2.2.2 Proposed framework for producing word sense association networks 17

2.2.3 Conversion of word sense association network from intermediate form to final form 20

2.3 Discussion 22

2.3.1 Potential impact of word sense association networks on natural language processing tasks 22

2.3.2 Linkage of word sense association network with WordNet 23

2.4 Summary 25

Chapter 3 Extraction of Translation Equivalents Based on Contextual Similarity **26**

3.1 Goal and approach 26

3.2 Basic idea 27

3.2.1 Similarity of contexts in two languages 27

3.2.2 Types of co-occurrence 30
3.2.3 Handling pairs of corresponding documents separately 31
3.3 Algorithm 33
3.3.1 Outline 33
3.3.2 Extraction of words 34
3.3.3 Extraction of word co-occurrence 35
3.3.4 Calculation of similarity between words 35
3.3.5 Selection of pairs with mutually highest similarity 38
3.3.6 Feedback of extracted pairs of translation equivalents 39
3.4 Experimental evaluation 39
3.4.1 Method and materials 39
3.4.2 Characteristics of patent specification documents 40
3.4.3 Results 42
3.5 Discussion 44
3.5.1 Advantages of proposed method 44
3.5.2 Performance compared with previous methods 44
3.5.3 Limitations and directions for extension 46
3.6 Related work 46
3.7 Summary 48

Chapter 4 Iterative Calculation of Sense-vs.-Clue Correlations Based on Translingual Alignment of Word Associations 50

4.1 Goal and approach 50
4.2 Basic idea 52
4.2.1 Translingual alignment of word associations 52
4.2.2 Coping with alignment ambiguity 53
4.2.3 Coping with alignment failure 56
4.3 Algorithm 58
4.3.1 Outline 58
4.3.2 Definition of word senses 59
4.3.3 Extraction of word associations 60
4.3.4 Translingual alignment of word associations 61
4.3.5 Calculation of correlations between senses and clues 62
4.3.6 Example of convergence of sense-vs.-clue correlations 63
4.4 Word sense disambiguation using sense-vs.-clue correlation matrix 65
4.5 Experimental evaluation 66
4.5.1 Method and materials 66

4.5.2 Comparison of formulae for defining second component plausibility factor 69
4.5.3 Comparison of formulae for defining sense score 71
4.5.4 Detailed evaluation of word sense disambiguation results 72
4.5.5 Comparison with alternative methods 75
4.5.6 Sensitivity to bilingual dictionary and corpus 77
4.6 Discussion 79
4.6.1 Strong and weak points of proposed method 79
4.6.2 Limitations and directions for extension 80
4.7 Related work 82
4.8 Summary 84

Chapter 5 Clustering of Translation Equivalents Based on Similarity of Translingual Distribution Patterns 86
5.1 Goal and approach 86
5.2 Basic idea 87
5.2.1 Clustering translation equivalents of target word 87
5.2.2 Translingual distributional word clustering 88
5.2.3 Similarity based on distribution pattern and subordinate distribution pattern 89
5.3 Algorithm 91
5.3.1 Outline 91
5.3.2 Calculation of sense-vs.-clue correlation matrices 92
5.3.3 Calculation of sense similarity matrix 93
5.3.4 Merging similar senses 93
5.3.5 Illustrative example of clustering 94
5.3.6 Variations 96
5.4 Experimental evaluation 97
5.4.1 Method and materials 98
5.4.2 Evaluative measures 98
5.4.3 Comparison between variations 100
5.4.4 Detailed analysis of example results 103
5.4.5 Comparison with alternative methods 105
5.5 Discussion 105
5.5.1 Advantages of proposed method 105
5.5.2 How to evaluate results of word sense acquisition 108
5.5.3 Limitations and directions for extension 108

5.6 Related work 109
5.7 Summary 110

Chapter 6 Conclusion **112**
6.1 This work 112
6.2 Future work 113

Acknowledgments **114**

References **116**

Publication List **124**

List of Figures

2.1 Example word sense definitions 12
2.2 Word association network and word sense association network 14
2.3 Proposed framework for producing word sense association network 18
2.4 Intermediate form of word sense association network 19
2.5 Linkage of word sense association network with WordNet 24

3.1 Outline of proposed method for extracting translation equivalents 27
3.2 Similarity between sets of co-occurring words along with co-occurrence frequencies 29
3.3 Types of co-occurrence 31
3.4 Conflict between pairs of translation equivalents 32
3.5 Flow of translation equivalent extraction 34
3.6 Example of overestimated intersection of sets of co-occurring words 37

4.1 Outline of method proposed for calculating correlations between senses and clues 51
4.2 Example sets of translingually alignable accompanying words 54
4.3 Example set of accompanying words regardless of translingual alignability 57
4.4 Flow of sense-vs.-clue correlation calculation 59
4.5 Convergence of sense-vs.-clue correlations 64
4.6 Flow of word sense disambiguation using sense-vs.-clue correlation matrix 65
4.7 Examples of sense determination 67
4.8 WSD performance using sense-vs.-clue correlation matrices produced by Method I 69
4.9 WSD performance using sense-vs.-clue correlation matrices produced by Method II 70
4.10 WSD performance using different formulae for defining the score 71
4.11 Distribution of *F*-measures: proposed method vs. baseline method 74
4.12 WSD performance using sense-vs.-clue correlation matrices produced by Method II and alternatives 1 and 2 76
4.13 WSD performance using sense-vs.-clue correlation matrices produced using reduced dictionaries 78

4.14 Results of WSD using Wall Street Journal and Mainichi Shimbun corpora 79

5.1 Outline of proposed method for clustering translation equivalents 87
5.2 Distribution pattern vs. subordinate distribution pattern 90
5.3 Flow of translation equivalent clustering 92
5.4 Clustering translation equivalents of "promotion" 95
5.5 Recall of senses and accuracy of sense definitions for ten variations of proposed translation equivalent clustering method 102
5.6 Clustering results for four target words 104
5.7 Clustering results using alternative bilingual dictionaries to select translation equivalents for two target words 106
5.8 Clustering with proposed method and two alternative methods 107

List of Tables

3.1 Profile of patent specification documents used for evaluation 41
3.2 Recall and precision of translation equivalent extraction 43
3.3 Methods for extracting translation equivalents from comparable corpora 48

4.1 Examples of English word association and their Japanese counterparts
along with mutual information values 56
4.2 Results of word sense disambiguation for six test words 73

5.1 Evaluation of word sense acquisition for "promotion" 100

Chapter 1

Introduction

1.1 Motivation

Despite over half a century of research and development, natural language processing (NLP) technologies are still immature, and there are few practical applications. Although machine translation systems have been commercialized, they are of limited use in assimilating information written in foreign languages, and they are far from replacing human translators. While information retrieval is a relatively well-established discipline, with many practical systems in use, the application of NLP techniques to information retrieval is in its infancy. Lower level processing, such as morphological analysis, plays a role in information retrieval, but the usefulness of syntactic and semantic processing for information retrieval has yet to be fully accepted. Other NLP applications, including information extraction, text summarization, and question answering, remain in the laboratory stage.

NLP technologies are not yet fully practical primarily because of the open-endedness of application systems and the idiosyncrasies of linguistic phenomena. "Open-ended" means that both vocabulary and sentence structure cannot be restricted even if the subject domain and/or the users are restricted. "Idiosyncrasy" is a distinguishing characteristic of natural language. Although the core part of a language can be formalized using a rather small number of rules, there are many exceptions specific to a particular word. Therefore, real-world systems must have an extensive range of linguistic knowledge.

NLP systems generally consist of processing engines and knowledge components such as grammars and lexicons. Separation of the knowledge components from the processing engines is one way to cope with the open-endedness and the idiosyncrasies. However, the development of knowledge components, which is labor-intensive, is still a bottleneck. From the practical application viewpoint, reducing the development cost of the knowledge components is crucial.

The 1990s witnessed a paradigm shift in the field of NLP: from a rationalist approach to an empiricist approach (Church and Mercer 1993). The availability of a growing amount of textual data in electronic form, together with advances in computing

power, has enabled us to take a corpus-based approach, i.e., automatic knowledge acquisition from corpora. Automatic knowledge acquisition is attractive because it is self-contained. That is, NLP techniques and statistical methods can be used to acquire linguistic knowledge from text corpora; feedback of the acquired knowledge strengthens the NLP techniques, which in turn accelerates the acquisition of linguistic knowledge.

Automatic knowledge acquisition has the potential of overcoming the problems of open-endedness and idiosyncrasy. It will enable the knowledge components of NLP systems to cover the range of linguistic phenomena observed in a given domain at reduced cost. Furthermore, it will improve the portability of NLP techniques to new domains as well as to new problems.

1.2 Purpose of study

This thesis addresses the acquisition of lexico-semantic knowledge, specifically the knowledge of the polysemy and synonymy of words, from corpora. Polysemy and synonymy are distinguishing characteristics of natural language. That is, a word often has two or more meanings, and two or more words often have (nearly) the same meaning. These have been major bottlenecks restricting machine understanding of natural language.

Word sense disambiguation (WSD) and synonym identification are key issues in most NLP tasks. Machine translation is a representative task in which WSD plays a central role. WSD is indispensable for assigning an appropriate word in the target language to an input word in the source language. In information retrieval, synonym identification is essential for improving recall, and WSD is essential for improving precision. WSD is also important in connection with query expansion, which is a means of improving retrieval effectiveness; without WSD, query expansion often results in degraded precision.

In this thesis, I describe the innovative approach that I developed with the assistance of my colleagues in Hitachi's Central Research Laboratory for acquiring knowledge of word senses from a bilingual comparable corpus and a bilingual dictionary. A novel semantic lexicon, a "word sense association network," is produced that enables effective WSD and synonym identification. This network can serve as a basis for semantics-oriented natural language processing.

Use of bilingual resources is a distinguishing feature of our approach. The underlying assumption is that two languages are more informative than one. That is, contrastive treatment of two languages enables us to learn much about each language. Polysemy and synonymy is idiosyncratic to each language, and they are not parallel

between languages, especially between languages with different origins, like English and Japanese. While non-parallelism between languages is a major obstacle to machine translation and multilingual NLP, it enables word senses in different languages to be acquired more easily.

Bilingual corpora can be divided into parallel corpora and comparable corpora. Parallel corpora are those consisting of two language texts that are translations of one another. Comparable corpora are those consisting of two language texts that are comparable to one another. There is a wide range of comparability; one extreme is pairs of articles in different languages describing the same event or idea, and the other extreme is a combination of corpora in different languages for the same domain, e.g., a combination of a Wall Street Journal corpus and a Nihon Keizai Shimbun corpus. A great deal of work has been done on knowledge acquisition from parallel corpora. In contrast, there has been a very little work on knowledge acquisition from comparable corpora, which is much more difficult because of uncertain correspondence between the language texts. However, given the extremely limited availability of large-scale parallel corpora, it is very important to develop techniques applicable to comparable corpora. Therefore, we aimed at producing word sense association networks from weakly comparable corpora.

The use of a bilingual dictionary as a knowledge source should be noted because it is a secondary language resource. Creating a word sense association network from only a bilingual corpus would be ideal, and doing so is possible in principle. That is, our approach could be combined with one for creating a bilingual dictionary from a bilingual corpus. However, given that manually compiled bilingual dictionaries are readily available, we have taken the realistic approach of using a bilingual dictionary as an auxiliary knowledge source. Existing bilingual dictionaries, of course, are not complete, so we have also developed a method for augmenting a bilingual dictionary using bilingual corpora. Using an augmented bilingual dictionary will result in improved word sense association networks.

It should also be mentioned that we do not use an existing semantic lexicon in our approach. Using a lexicon that already defines word senses would make the task easier. However, the coverage of such lexicons is incomplete. In particular, domain-specific senses are often missing, while many irrelevant or rare senses are included. Therefore, it is essential to divide and define the senses of each word corpus-dependently. Senses given by an existing lexicon could be used as seeds, if they were defined formally. However, most lexicons define word senses with descriptive texts, which prevents them being used as seed lexicons.

To sum up, the primary purpose of this study was to develop a method for producing a word sense association network from a bilingual comparable corpus and a bilingual dictionary. Its secondary purpose was to develop a method for extracting translation equivalents from a bilingual corpus. These methods are applicable to any language pair, although we implemented and evaluated them for English and Japanese.

1.3 Survey of previous work in related areas

1.3.1 Semantic lexicons

A number of large-scale semantic lexicons have already been developed. Most of them, including WordNet and the EDR concept dictionary, have been handcrafted, while a few, like MindNet, have been produced automatically.

WordNet, which was developed at Princeton University, is a machine-readable semantic lexicon of English based on psycholinguistic theories (Miller 1990; Fellbaum 1998). It structures lexical information in terms of word senses and encompasses a majority of nouns, verbs, adjectives, and adverbs. Words of the same part of speech that can be used to express the same meaning are grouped into a set of synonyms, called a "synset." Concepts represented as synsets are further connected through a small set of lexico-semantic relations. The dominant relation is hypernymy/hyponymy (i.e., is-a links). Meronym relations (i.e., part-of links) and antonymy relations are also encoded. WordNet version 1.6 contains 121,962 words and 99,643 synsets (concepts).

WordNet has been applied to many NLP tasks including word sense disambiguation, machine translation, conceptual indexing, information retrieval, and text classification, and its usefulness has been proved along with its limitations. A major difficulty in using WordNet in a specific domain is that many of the domain-specific concepts and relations are missing, while many irrelevant ones are present (Ioannis-Dimitrios and Dimitris 2002). WordNet also lacks links between topically related concepts (Harabagiu, et al. 1999; Agirre, et al. 2001).

The success of WordNet led to the emergence of the EuroWordNet project (Vossen 1998). The objective of EuroWordNet is to develop an extensive multilingual lexicon with networks for the French, German, Spanish, Dutch, Italian, Czech, Estonian, and English languages. The project also aims at a language-independent set of semantic concepts linking the language networks together.

The EDR dictionaries, which were developed at the Japan Electronic Dictionary Research Institute, Ltd., include a concept dictionary of the Japanese and English languages (EDR 1990a; Yokoi 1995). It structures 400,000 concepts, each of which is defined descriptively, based on hypernymy/hyponymy relations. In addition, case

relations such as agent, object, manner, and implement are encoded as well as synonym and meronym relations.

MindNet, developed at Microsoft Research, has a distinguishing characteristic—it was produced automatically from two machine-readable dictionaries: the Longman Dictionary of Contemporary English (LDOCE) and the American Heritage Dictionary, 3rd Edition (Richardson, et al. 1998). Its basic units are word senses defined in the source dictionaries, and different types of semantic relations between them are extracted by parsing the sense definitions and example sentences. For example, LDOCE (1995) defines one sense of "car" as

> "a vehicle with 3 or usu. 4 wheels and driven by a motor, esp. one for carrying people."

Parsing this definition sentence results in the following relations:

> "car" == hypernym ==> "vehicle"
> "car" == part ==> "wheel"
> "car" == purpose ==> "carry"
> "drive" == object ==> "car"
> "drive" == means ==> "motor"
> "carry" == object ==> "people"

Word sense disambiguation is done during parsing, so relations between word senses, not between words, are extracted. MindNet consists of 713,000 relations extracted from 191,000 sense definitions and 58,000 example sentences. However, it still provides spotty coverage of the English language, which is a limitation common to lexicons produced from machine-readable dictionaries.

1.3.2 Translation equivalent extraction

A bilingual dictionary is a key component of machine translation systems—its quality affects the quality of the translated sentences. Its coverage is also important; it should cover not only the 10,000 or so general words but also terminology in a specific application domain. Therefore, automatic extraction of translation equivalents from bilingual corpora has been one of the major topics since the beginning of corpus-based NLP research, and a number of practical methods have been developed. Methods for automatically extracting translation equivalents from bilingual corpora can be divided into statistical ones and linguistic ones.

The statistical methods assess pair-wise correlations between words in different languages based on their distribution in a parallel corpus (Gale and Church 1991a; Kupiec 1993; Dagan, et al. 1993; Inoue and Nogaito 1993; Fung 1995; Kitamura and

Matsumoto 1996; Melamed 1997b). They, with some exceptions like Fung's (1995) method, presuppose that the input parallel corpus is aligned sentence by sentence. For example, Gale and Church's (1991a) method measures the correlation between words in the first and second languages using χ^2-like statistics based on a two-by-two contingency table showing the number of aligned sentences containing both words, of those containing the first-language word but not the second-language word and vice versa, and of those containing neither word.

Automatic methods for aligning sentences in a parallel corpus, which are prerequisite to most of the statistical translation equivalent extraction methods, have also been developed (Brown, et al. 1991a; Gale and Church 1991b; Kay and Roscheisen 1993; Chen 1993; Melamed 1997a). These methods assume that sentence order is maintained, even though a sentence may correspond to two or more consecutive sentences and another sentence may have no counterpart. Correspondences between sentences are established based on length correlation (the tendency of short sentences to translate into short sentences, and long into long) and/or lexical anchors (e.g., dictionary-based word pairs or cognate pairs).

The linguistic methods extract pairs of compound words that are translations of one another by examining the correspondence between their constituent words, with the assistance of a bilingual dictionary (Yamamoto and Sakamoto 1993; Ishimoto and Nagao 1994). They aim mainly at constructing bilingual dictionaries of domain-specific technical terms, most of which are compound words. However, their usefulness is limited, since the constituent words of equivalent terms do not always have a one-to-one correspondence. A hybrid approach combining a linguistic method with a statistical method has also been proposed (Kumano and Hirakawa 1994).

1.3.3 Corpus-based word sense disambiguation

Word sense disambiguation is the process of assigning the appropriate sense to a word in a given context. It is an intermediate task necessary for accomplishing most NLP tasks, especially machine translation and information retrieval, and it is often cited as one of the most important problems in NLP research today.

WSD has been of interest and a concern since the earliest days of machine translation research in the 1950s. The earliest work on data-driven WSD, or corpus-based WSD, was seen in the 1970s. Weiss (1973) and Kelley and Stone (1975) demonstrated that disambiguation rules can be learned from a manually sense-tagged corpus. However, development of full-fledged automatic WSD methods had to wait until large amounts of machine-readable text became available. In the late 1980s, Black

(1988) developed a decision tree model using a corpus of 22 million words, after manually sense-tagging 2,000 concordance lines for 5 target words. Since then, supervised learning from sense-tagged corpora has been used by many researchers (Zernik 1991; Hearst 1991; Leacock, et al. 1993; Gale, et al. 1993; Bruce and Wiebe 1994; Miller et al. 1994; Niwa and Nitta 1994). However, supervised WSD is costly because it requires manually tagging the sense onto each instance of a polysemous word in a training corpus.

A number of bootstrapping methods have been developed to reduce the sense-tagging cost. Hearst (1991) proposed an algorithm called CatchWord that includes a training phase during which each occurrence of a word to be disambiguated is manually sense-tagged in ten or so occurrences. Statistical information extracted from the contexts of these occurrences is then used to disambiguate other occurrences. If another occurrence can be disambiguated with certitude, additional statistical information is extracted from the context of the newly disambiguated occurrence. Thus, the knowledge useful for disambiguation is increased incrementally. Basili (1997) proposed a class-based bootstrapping method for semantic tagging in specific domains.

A variety of unsupervised WSD methods that use a machine-readable dictionary or thesaurus in addition to a corpus have also been proposed (Yarowsky 1992; Luk 1995; Yarowsky 1995; Karov and Edelman 1998). For example, Yarowsky (1995) proposed an unsupervised method that uses the textual definitions of senses as seeds. First, it identifies seed collocations representative of each sense and tags all training examples containing the collocations with the seed's sense label. Then it identifies other collocations that reliably partition the seed training examples, and the resulting classifier is used to tag more training examples. This procedure is repeated until no more training examples can be tagged. Yarowsky reported that this unsupervised method achieved nearly the same performance as a supervised method.

Use of bilingual corpora is another way to avoid manually sense-tagging training data (Brown, et al. 1991b; Gale, et al. 1992b). The underlying assumption is that different senses of a given word often translate differently into another language. A parallel corpus can be aligned word for word automatically (Gale and Church 1991a; Dagan, et al. 1993). In the resulting aligned corpus, the senses of words in the text of one language are indicated by their counterparts in the text of the other language. Thus, costly manual sense-tagging can be avoided. This method has some limitations due to the limited availability of large-scale parallel corpora. To solve this, Dagan and Itai (1994) proposed a method that uses a second-language monolingual corpus and a bilingual dictionary instead of a parallel corpus.

While we have surveyed the major approaches to corpus-based WSD, we have not compared their performances. This is because they were evaluated using their own rather small sets of typical polysemous words along with their own training and test corpora and sense inventories, making it difficult to compare their performances. A series of evaluation exercises has recently been developed so that different methods can be evaluated on a common task basis (Kilgarriff and Rosenzweig 2000; SENSEVAL-2 2001), with the result that research on WSD is advancing from the feasibility demonstration stage to the large-scale evaluation and improvement stage.

1.3.4 Distributional word clustering

Clustering of words based on their semantic similarity is increasingly being used in a number of natural language processing tasks. That is, word clustering addresses the problem of data sparseness in statistical language processing; it enables generalization of statistical language models, particularly models for deciding among alternative analyses proposed by a grammar (Brill, et al. 1990; Brown, et al. 1992; Clark and Weir 2000). Word clustering is also useful in information retrieval; query expansion using similar words improves the effectiveness of information retrieval.

Semantic similarity between words can be calculated based on either the taxonomical relationship, such as hyponymy and synonymy, between the words or the distributional characteristics of the words. While it is almost impossible to exhaustively extract pairs of words having a taxonomical relationship from corpora, a fraction of them can be extracted (Hearst 1992). In contrast, the distributional characteristic can be calculated for every word in a corpus. With the growing availability of large text corpora, a great deal of work on distributional word clustering has been done over the last dozen or so years (Hindle 1990; Ruge 1991; Pereira, et al. 1993; Grefenstette 1994; Ushioda 1996; Li and Abe 1998; Lin 1998; Allegrini, et al. 2000; Lin and Pantel 2002).

Distributional word clustering is based on the "distributional hypothesis," that is, words occurring in similar contexts tend to be similar (Harris 1985). Each word is characterized by a vector or a weighted set consisting of words that co-occur with it, and the similarity between words is defined as the similarity between their vectors or weighted sets. Methods proposed so far differ mainly in which type of co-occurrence and which similarity metric are used. For example, Hindle's (1990) method classifies nouns by using co-occurrence in subject-verb and verb-object relations and a similarity metric based on mutual information. The method of Pereira, et al. (1993) classifies nouns according to their distribution as direct objects of verbs. It uses a measure of distributional dissimilarity rather than similarity; dissimilarity between two nouns is

defined as the relative entropy (Kullback-Leiber distance) of the corresponding conditional verb distributions.

1.4 Thesis overview

Chapter 2 gives an overview of our proposed approach. That is, the word sense association network is described, and a framework for producing it from a bilingual comparable corpus and a bilingual dictionary is presented.

Chapter 3 presents a method for extracting pairs of translation equivalents based on contextual similarity from bilingual corpora that are not aligned sentence by sentence. The method is evaluated through an experiment using pairs of Japanese and English patent-specification documents.

Chapter 4 describes the key technique of our approach—an iterative algorithm for calculating correlations between senses of a polysemous word and clues identifying its sense based on translingual alignment of word associations. It is evaluated through a WSD experiment using Wall Street Journal and Nihon Keizai Shimbun corpora and the EDR bilingual dictionary.

Chapter 5 describes our word sense acquisition method, i.e., clustering translation equivalents of a polysemous word based on their translingually aligned distribution patterns. Its effectiveness is demonstrated through an experiment using the same corpora and dictionary as in the experiment of Chapter 4.

Chapter 6 summarizes the results of the study and suggests directions for future work.

Chapter 2

From Word Associations to Word Sense Associations: An Approach Using Bilingual Corpora

2.1 What is a word sense association network?

2.1.1 Word association: its usefulness and limitations

The sense of a word is suggested by the company it keeps. For instance, on the one hand, we find the English word "race" in the company of "black," "gender," "Hispanic," "minority," etc., and these co-occurring words suggest the sense of "race" to be "a group of people." On the other hand, "race" is also found in the company of "car," "circuit," "gamble," "winner," etc., and these co-occurring words suggest the sense of "race" to be "a competition."

The growing availability of text corpora enables word associations, i.e., pairs of significantly related words, to be collected using statistical methods (Church and Hanks 1990; Calzolari and Bindi 1990; Smadja 1993). Typically, pairs of words co-occurring with each other are extracted from a corpus, the correlation such as mutual information is calculated for each pair, and pairs with a correlation larger than a threshold are selected.

Word associations are useful for solving some problems in NLP. For example, they provide clues for parsing ambiguous structures such as nominal compounds, coordinated structures, and prepositional-phrase attachments (Hindle and Rooth 1993). They also provide constraints for resolving pronoun references (Dagan and Itai 1990). Furthermore, they constrain the language models for speech recognition and optical character recognition.

Word associations are potentially useful for solving other problems, including word sense disambiguation and synonym identification, since clues for treating senses are implicit. People use real-world knowledge and common sense to recognize the associated senses when a word association is given. That is, they can fairly well identify which specific senses of the words are relevant to the association. In contrast, machines, which lack real-world knowledge and common sense, cannot do this unless word sense associations, not word associations, are given explicitly. Word associations also provide

clues for identifying synonyms because synonyms are associated with similar sets of words. However, synonymy relations are only represented implicitly, so the usefulness of word associations is yet to be determined. Word associations should thus be converted into word sense associations to enable machines to recognize word senses.

2.1.2 Definition of word senses

From the engineering standpoint, we presume several different senses can be recognized for a polysemous word.[1] We do not extend the philosophical discussion on word sense (Kilgarriff 1998). Instead, our focus is on how to define word senses. We have the following alternatives (example word sense definitions are shown in Fig. 2.1).

(1) Textual definition

Textual definitions, which are commonly given in monolingual dictionaries for use by people, have been used in a great deal of work on WSD because they contain clue words for determining the sense (Lesk 1986; Veronis and Ide 1990; Guthrie, et al. 1991; Cowie, et al. 1992; Dolan 1994; Luk 1995; Yarowsky 1995; Karov and Edelman 1998). However, they have no additional advantage, and they are difficult for machines not only to understand but also to generate. They are thus not suitable for our purpose.

(2) Thesaurus category

Thesaurus categories are sometimes used for sense definitions (Walker and Amsler 1986; Yarowsky 1992). That is, labeling a word with a thesaurus category code corresponds to defining the sense of the word, assuming that different senses of a word belong to different categories. However, category codes are relatively weak as sense definitions because they only define senses indirectly.

(3) Synonyms in the same language

Using synonyms is an effective way to define word senses. In particular, the WordNet synsets (Miller 1990; Fellbaum 1998) have been widely used in work on WSD (Li, et al. 1995; Agirre and Rigau 1996; Mihalcea and Moldovan 1999; Kilgarriff 2001). However, rather than use such manually prepared sense definitions, we need to produce sense definitions from a corpus. Therefore, whether senses can be defined using synonyms depends on how accurately all the words in the language can be clustered into subsets consisting of synonyms.

(4) Translation equivalents in another language

---

[1] We do not distinguish between polysemous words and homographs in this thesis. While a "polysemous word" is a word that has two or more interrelated but different meanings, "homographs" are different words that are the same in spelling but different in meaning. The distinction between them, however, is not always clear. We express both as polysemous words and treat them uniformly.

race

- one of the main groups that humans can be divided into according to the color of their skin and other physical features
- a competition in which each competitor tries to run, drive, etc. fastest and finish first

Longman Dictionary of Contemporary English (LDOCE 1995)

(a) Textual definition

● CLASS 4: Matter	● CLASS 7: Volition
III. Organic matter	V. Support and opposition
E. Mankind	D. Contention
418. People	796. Contention
race	race

Roget's International Thesaurus (ROGET 1977)

(b) Thesaurus category

race

- people, stock, ethno-, ethnic group, community, nationality, nation
- contest, derby, heat, lap, footrace, competition, rivalry, vying

(c) Synonyms in the same language

race

- 人種<*JINSHU*>, 民族<*MINZOKU*>, 国民<*KOKUMIN*>, 種族<*SHUZOKU*>
- 競走<*KYOUSOU*>, レース<*REESU*>, 競馬<*KEIBA*>, 競争<*KYOUSOU*>, 戦い<*TATAKAI*>

(d) Translation equivalents in another language

Figure 2.1 Example word sense definitions

Use of translation equivalents is another potentially effective way to define word senses. It serves as the basis for WSD using a bilingual corpus (Brown, et al. 1991b; Gale, et al. 1992b) and for WSD using a second-language monolingual corpus (Dagan and Itai 1994). These methods treat translation equivalents of a target word[2] as if they corresponded to different senses of the target word. Strictly, however, it is necessary to

[2] We use "target word" in this thesis to indicate the word whose senses are to be defined, disambiguated, or acquired. This definition of "target" differs from that normally used in machine translation, i.e., translation of a word in the source language into a target language.

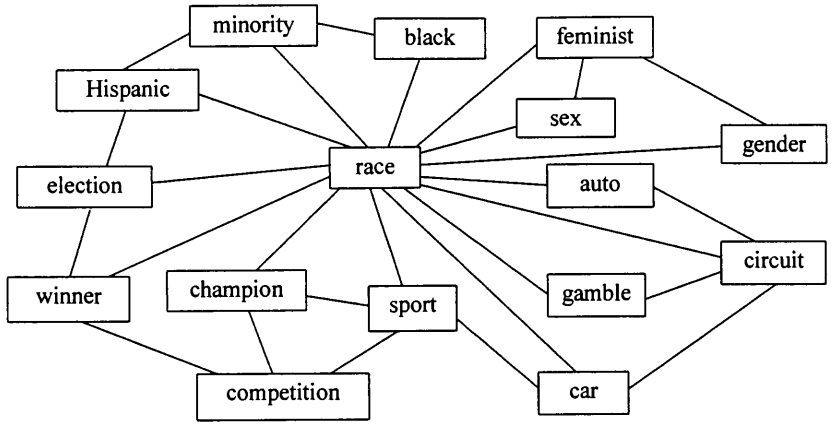
group synonymous translation equivalents. Therefore, whether senses can be defined using translation equivalents depends on how accurately a set of translation equivalents of a target word can be clustered into subsets consisting of synonymous ones. Translation equivalents are useless if they preserve the ambiguity of the target words, a serious problem for pairs of languages with the same origin, like English and French. However, it is less serious for pairs of languages with different origins, like English and Japanese.

It should be noted that defining senses using synonyms and using translation equivalents are almost the same in our approach using bilingual corpora. That is, a set of translation equivalents that defines a sense of a target word in a language can be regarded as a set of synonyms that defines a sense of another target word in another language. For example, a set of Japanese translation equivalents {レース<*REESU*>, 競争<*KYOUSOU*>, 競馬<*KEIBA*>} that defines the "competition" sense of the English word "race" can be regarded as a set of Japanese synonyms that defines the "race or competition" sense of the Japanese word "レース<*REESU*>."

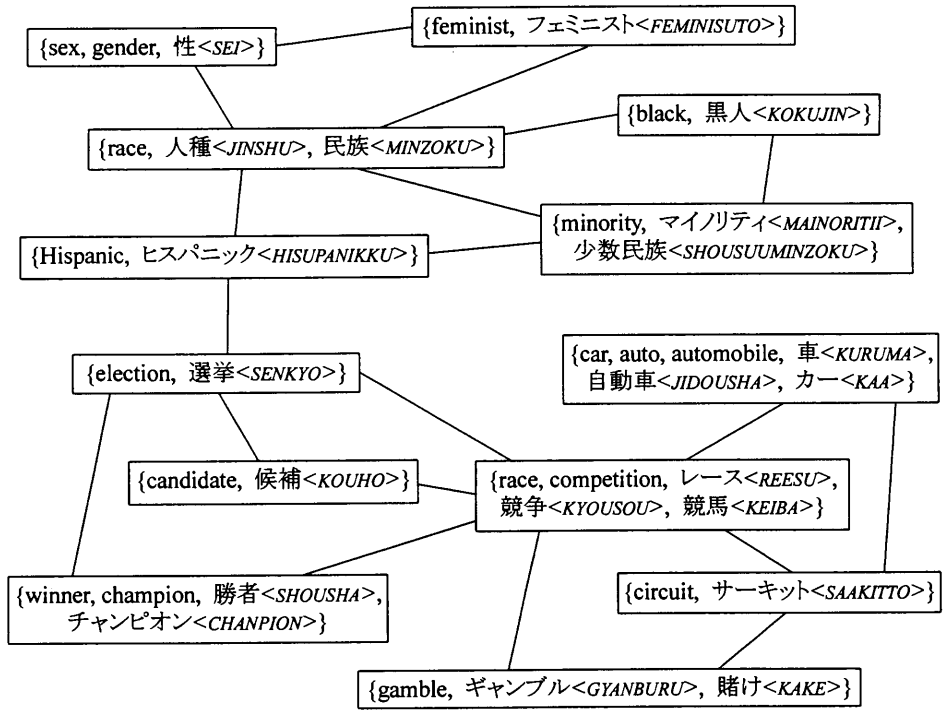
Thus, our choice is between defining senses using synonyms and defining senses using translation equivalents. The criterion is which is preferable: the clustering of all words in the language or the clustering of a set of translation equivalents for each target word. Obviously, the latter is computationally more efficient. In addition, it possibly produces better results because it clusters a restricted set of words corresponding to a particular target word. Therefore, we define senses using translation equivalents. That is, each sense of an English word is defined using a set of Japanese synonymous translation equivalents, and each sense of a Japanese word is defined using a set of English synonymous translation equivalents. We further unify the senses of the English words with those of the Japanese words. As a result, senses are defined using sets of bilingual synonyms.

2.1.3 Word sense association network

A word sense association network is a graph whose nodes represent word senses and edges represent associations between word senses. Each sense is defined as a set of bilingual synonyms. We call a similar graph whose nodes represent words and edges represent associations between words a word association network. Example fragments of a word association network and a word sense association network are shown in Fig. 2.2.



(a) Word association network



(b) Word sense association network

Figure 2.2 Word association network and word sense association network

In the word association network, a node has multiple senses, each of which is relevant to a subset of associations represented by edges connected to the node. However, the relevancy is not represented explicitly. In the example shown in Fig. 2.2(a), the node "race" has both the "group of people" sense and the "competition" sense. The former is relevant to associations with "black," "gender," etc., and the latter

is relevant to associations with "car," "circuit," etc. However, the relevancy is not explicit in the network.

In contrast, a node in the word sense association network has one sense that is relevant to all associations represented by edges connected to the node. The example in Fig. 2.2(b) includes two nodes that represent the two senses of the English word "race." The node labeled {race, 人種<*JINSHU*>, 民族<*MINZOKU*>} represents the "group of people" sense, and the one labeled {race, competition, レース<*REESU*>, 競争<*KYOUSOU*>, 競馬<*KEIBA*>} represents the "competition" sense. These nodes are connected to those representing associated word senses. That is, on the one hand, {race, 人種<*JINSHU*>, 民族<*MINZOKU*>} is connected to {black, 黒人<*KOKUJIN*>}, {sex, gender, 性<*SEI*>}, etc. On the other hand, {race, competition, レース<*REESU*>, 競争<*KYOUSOU*>, 競馬<*KEIBA*>} is connected to {car, auto, automobile, 車<*KURUMA*>, 自動車<*JIDOUSHA*>, カー<*KAA*>}, {circuit, サーキット<*SAAKITTO*>}, etc.

The word sense association network is useful not only as an internal knowledge component of NLP systems, but also as a human-machine interface. It enables word senses or concepts to be shared by humans and machines. This capability is very important because many NLP applications involve human-machine interaction. The word sense association network provides two ways of concept sharing—one is through sets of bilingual synonyms and the other is through sets of associated words.

It is obvious that bilingual people can recognize a sense defined as a set of bilingual synonyms. Unfortunately, however, monolingual people often encounter difficulty in recognizing a sense defined as a set of bilingual synonyms. For example, deleting the Japanese words, which English monolingual people cannot read, from {race, 人種<*JINSHU*>, 民族<*MINZOKU*>} and {race, competition, レース<*REESU*>, 競争<*KYOUSOU*>, 競馬<*KEIBA*>} results in {race} and {race, competition}, respectively. From {race, competition}, English monolingual people can still recognize the "competition" sense of "race." However, from {race}, they cannot recognize the "group of people" sense of "race." As this example shows, senses defined as a set of bilingual synonyms are not always effective for monolingual people.

In contrast, the set of associated words enables a sense to be always recognized by monolingual people. For example, English monolingual people recognize the "group of people" sense through the set of its associated words consisting of "black," "sex," "gender," "minority," etc., while Japanese monolingual people recognize the same sense through the set of its associated words consisting of "黒人<*KOKUJIN*>," "性<*SEI*>," "マイノリティ<*MAINORITII*>," "少数民族<*SHOUSUU-MINZOKU*>," etc.

Finally, we mention the similarity and dissimilarity of the word sense association

network to existing semantic lexicons. Both WordNet (Miller 1990; Fellbaum 1998) and the word sense association network define word senses by using sets of synonyms, although one is monolingual and the other is bilingual. In contrast, the EDR concept dictionary (EDR 1990a; Yokoi 1995) and MindNet (Richardson, et al. 1998) use textual definitions. Relations on which the word sense association network focuses are associations between word senses; it does not care which meanings the associations have. The word sense association network is thus dissimilar to the others, which focus on hypernymy/hyponymy, synonymy, meronymy, and case relations. Semantic lexicons that focus on word sense associations have yet to be developed because it is difficult to extract them exhaustively.

2.2 How word sense association networks are produced?

2.2.1 Basic idea

Our goal is to enable a word sense association network to be produced automatically from a bilingual comparable corpus and a bilingual dictionary. Using a conventional method, word association networks in respective languages can be produced from the respective language part of the corpus (Kaji, et al. 2000). Therefore, our problem is how to unify these word association networks into a word sense association network. Our basic idea is to align the word association networks of the two languages using a bilingual dictionary, through which both pair-wise word sense disambiguation and word sense acquisition are performed. The "pair-wise" word sense disambiguation means that a sense is assigned to a word based on each of the words associated with that word.

While similar ideas have been put forth, their purposes differ from ours and do not meet our need by themselves. Rapp (1995) proposed a method for permuting a word association matrix of one language to maximize the similarity to that of another language. Matrix permutation implies alignment of word associations. However, the method assumes one-to-one correspondence between words of the two languages; it never aligns a word with two or more words. Thus, it is unable to deal with multiple senses of a polysemous word, which is the central issue in producing a word sense association network.

Tanaka and Iwasaki (1996) proposed a method for finding a translation probability matrix that minimizes the distance between word association matrices of two languages. Their method deals with multiple translations, or senses, for a word. However, it does not track which associated word suggests which translation.

Dagan and Itai's (1994) WSD method using a second-language monolingual corpus and a bilingual dictionary is also closely related to our problem. It disambiguates

occurrences of polysemous words in first-language texts by using the second-language word association statistics. The disambiguation is based on "implicit" word sense associations suggested by second-language word associations; however, the method does not elicit word sense associations. In addition, the rationale for disambiguation is rather unreliable; for each first-language word association, the second-language one with the highest correlation is selected from a set of possible counterparts.

2.2.2 Proposed framework for producing word sense association networks

Our proposed framework for producing a word sense association network is illustrated in Fig. 2.3. It is divided into four modules: the translation equivalent extraction module, the sense-vs.-clue correlation calculation module, the translation equivalent clustering module, and the word sense unification module.

The translation equivalent extraction module improves the coverage of the bilingual dictionary, which affects the production of word sense association networks. It extracts pairs of translation equivalents from unaligned bilingual corpora based on contextual similarity. Translation equivalent extraction is described in detail in Chapter 3.

The other three modules unify the word association networks of the two languages into a word sense association network. The sense-vs.-clue correlation calculation module and the translation equivalent clustering module perform pair-wise word sense disambiguation and word sense acquisition, respectively. They depend on each other, and they are executed in an interleaving fashion. The interleaved execution of these two modules is the key to producing high-quality word sense association networks.

An intermediate form of a word sense association network consisting of sense inventories and sense-vs.-clue correlation matrices facilitates interaction between the two modules. A pair of a sense inventory and a sense-vs.-clue correlation matrix is created for each target word of both languages. The sense inventory lists the senses of the target word, each of which is defined as a set of the target word itself and its translation equivalents in the other language. The sense-vs.-clue correlation matrix represents correlations between the senses of the target word and the clues identifying the sense of the target word, i.e., words associated with the target word. An illustrative example of sense inventories and sense-vs.-clue correlation matrices for two English words, "race" and "gamble," and two Japanese words, "レース<*REESU*>" and "ギャンブル<*GYANBURU*>," is shown in Fig. 2.4.

The sense-vs.-clue correlation calculation module plays the most important role in the proposed framework. Given a sense inventory for a target word, it calculates correlations between the senses and the clues based on word association alignment

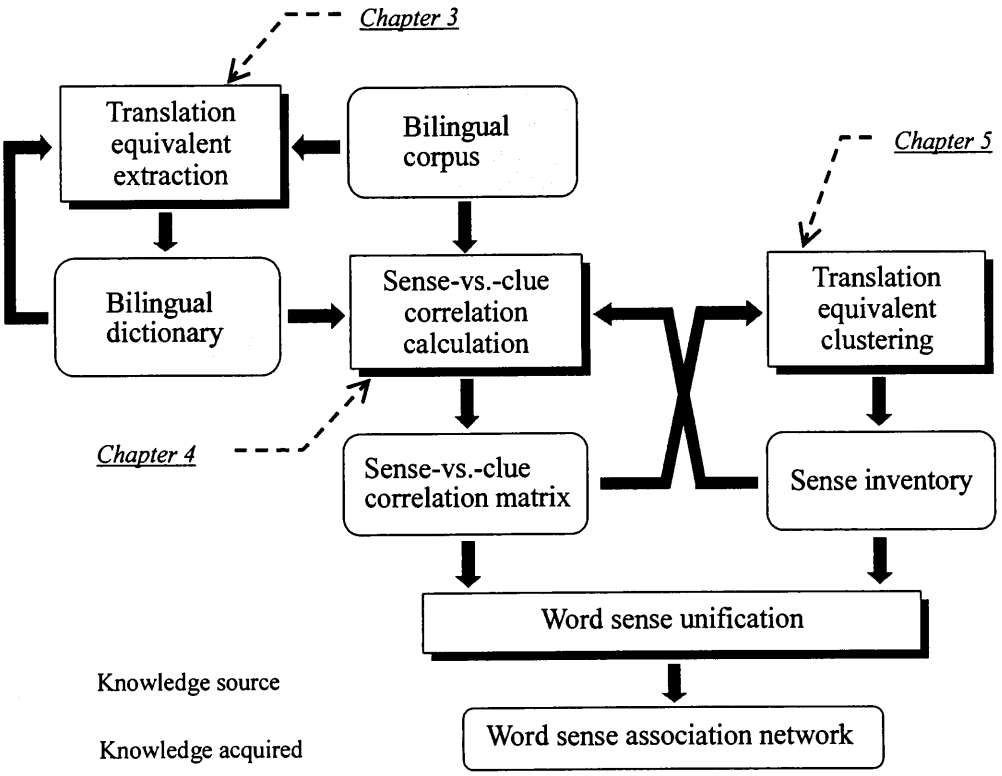


Figure 2.3 Proposed framework for producing word sense association network

using a bilingual dictionary. The method for calculating sense-vs.-clue correlations is described in detail in Chapter 4. The translation equivalent clustering module creates a sense inventory for a target word. It clusters translation equivalents of the target word based on the sense-vs.-clue correlation matrices produced by the sense-vs.-clue correlation calculation module. The method for clustering translation equivalents is described in detail in Chapter 5. The word sense unification module is an additional module that converts the word sense association network from the intermediate form to the final form. This module, whose function is straightforward, is described in the following subsection.

The framework described above is rather complicated. A possible simplified framework is one that serially executes the translation equivalent clustering module and the sense-vs.-clue correlation calculation module in that order. In this alternative, the translation equivalent clustering module cannot use information resulting from alignment of word associations, and, therefore, it suffers from not only the data sparseness problem but also the problems caused by corpus-irrelevant translation equivalents and corpus-irrelevant senses of polysemous translation equivalents. In

- RACE-1 = {race, 人種<*JINSHU*>, 民族<*MINZOKU*>}
- RACE-2 = {race, レース<*REESU*>, 競争<*KYOUSOU*>, 競馬<*KEIBA*>}

(a-1) Sense inventory for "race"

	candidate	gamble	gender	minority
RACE-1	1.30	0.10	3.58	1.50
RACE-2	2.15	1.04	0.25	0.07

(a-2) Sense-vs.-clue correlation matrix for "race"

- GAMBLE-1 = {gamble, 投機<*TOUKI*>, 冒険<*BOUKEN*>}
- GAMBLE-2 = {gamble, ギャンブル<*GYANBURU*>, 賭け<*KAKE*>}

(b-1) Sense inventory for "gamble"

	race	risk	stake	stock
GAMBLE-1	0.17	2.48	0.57	1.41
GAMBLE-2	1.04	0.35	1.34	0.56

(b-2) Sense-vs.-clue correlation matrix for "gamble"

- REESU-1 = {レース<*REESU*>, lace}
- REESU-2 = {レース<*REESU*>, race}

(c-1) Sense inventory for "レース<*REESU*>"

	織物 <*ORIMONO*>	絹 <*KINU*>	ギャンブル <*GYANBURU*>	ボート <*BOOTO*>
REESU-1	1.85	3.92	0.18	0.37
REESU-2	0.25	1.08	1.46	1.88

(c-2) Sense-vs.-clue correlation matrix for "レース<*REESU*>"

- GYANBURU-1 = {ギャンブル<*GYANBURU*>, gamble}

(d-1) Sense inventory for "ギャンブル<*GYANBURU*>"

	競馬 <*KEIBA*>	ボート <*BOOTO*>	マージャン <*MAAJAN*>	レース <*REESU*>
GYANBURU-1	1.25	2.50	2.12	1.46

(d-2) Sense-vs.-clue correlation matrix for "ギャンブル<*GYANBURU*>"

Figure 2.4 Intermediate form of word sense association network

contrast, the proposed framework avoids these difficulties, resulting in more accurate word sense acquisition (See Subsection 5.2.2 for details).

Finally, we mention the computational load of the proposed framework. The main modules, i.e., the sense-vs.-clue correlation calculation module and the translation

equivalent clustering module, are executed for each target word. The executions for the target words are independent. In each execution, a restricted set of words, i.e., the target word and its associated words together with their translation equivalents, are handled. Thus, the framework eases the burden of computation.

2.2.3 Conversion of word sense association network from intermediate form to final form

The intermediate form of the word sense association network is converted into the final form in two steps: monolingual unification of sense-clue pairs and translingual unification of word senses. While the monolingual unification is straightforward, the translingual unification is slightly problematic because the word sense associations in the two languages may conflict.

(1) Monolingual unification of sense-clue pairs

First, based on the assumption of one sense per clue, a sense-vs.-clue correlation matrix is converted into a set of sense-clue pairs. That is,

$$\{(\mathrm{S}(x,i),x') \mid C(S(x,i),x') = \max_{i'} C(S(x,i'),x')\},$$

where $S(x, i)$ denotes the i-th sense of target word x, x' denotes a clue, and $C(S(x, i), x')$ denotes the correlation between $S(x, i)$ and x'. Then, sense-clue pairs in which the target word and the clue are reversed are unified into a word sense association, i.e.,

$$(S(x_1, i_1), x_2) \text{ and } (\mathrm{S}(x_2, i_2), x_1) \Rightarrow (S(x_1, i_1), \mathrm{S}(x_2, i_2)).$$

Thus, word association (x_1, x_2) is converted into word sense association $(S(x_1, i_1), \mathrm{S}(x_2, i_2))$.

[Example]

The sense-vs.-clue correlation matrix for "race" in Fig. 2.4(a-2) results in the following sense-clue pairs:

({race, 人種<*JINSHU*>, 民族<*MINZOKU*>}, gender)

({race, 人種<*JINSHU*>, 民族<*MINZOKU*>}, minority)

({race, レース<*REESU*>, 競争<*KYOUSOU*>, 競馬<*KEIBA*>}, candidate)

({race, レース<*REESU*>, 競争<*KYOUSOU*>, 競馬<*KEIBA*>}, gamble)

Likewise, the sense-vs.-clue correlation matrix for "gamble" in Fig. 2.4(b-2) results in the following sense-clue pairs:

({gamble, 投機<*TOUKI*>, 冒険<*BOUKEN*>}, risk)

({gamble, 投機<*TOUKI*>, 冒険<*BOUKEN*>}, stock)

({gamble, ギャンブル<*GYANBURU*>, 賭け<*KAKE*>}, race)

({gamble, ギャンブル<*GYANBURU*>, 賭け<*KAKE*>}, stake)

As a result, word association (race, gamble) is converted into word sense association ({race, レース<*REESU*>, 競争<*KYOUSOU*>, 競馬<*KEIBA*>}, {gamble, ギャンブル<*GYANBURU*>, 賭け<*KAKE*>}).

Similarly, according to the sense-vs.-clue correlation matrices in Fig. 2.4(c-2) and (d-2), word association (レース<*REESU*>, ギャンブル<*GYANBURU*>) is converted into word sense association ({レース<*REESU*>, race}, {ギャンブル<*GYANBURU*>, gamble}).

(2) Translingual unification of word senses

The pairs of word sense associations in the two languages that correspond to each other are unified, i.e.,

$(S(x_1, i_1), S(x_2, i_2))$, and $(S(y_1, j_1), S(y_2, j_2)) \Rightarrow (S(x_1, i_1) \cup S(y_1, j_1), S(x_2, i_2) \cup S(y_2, j_2))$ if and only if

$$x_1 \in S(y_1, j_1),\ y_1 \in S(x_1, i_1),\ x_2 \in S(y_2, j_2), \text{ and } y_2 \in S(x_2, i_2).$$

[Example]

({race, レース<*REESU*>, 競争<*KYOUSOU*>, 競馬<*KEIBA*>}, {gamble, ギャンブル<*GYANBURU*>, 賭け<*KAKE*>}) and

({レース<*REESU*>, race}, {ギャンブル<*GYANBURU*>, gamble})

are unified into

({race, レース<*REESU*>, 競争<*KYOUSOU*>, 競馬<*KEIBA*>}, {gamble, ギャンブル<*GYANBURU*>, 賭け<*KAKE*>}),

since

race∈{レース<*REESU*>, race},

レース<*REESU*>∈{race, レース<*REESU*>, 競争<*KYOUSOU*>, 競馬<*KEIBA*>},

gamble∈{ギャンブル<*GYANBURU*>, gamble}, and

ギャンブル<*GYANBURU*>∈{gamble, ギャンブル<*GYANBURU*>, 賭け<*KAKE*>}.

Note that two or more pairs of word sense associations in the two languages often result in the same unified word sense association.

[Example]

({race, レース<*REESU*>, 競争<*KYOUSOU*>, 競馬<*KEIBA*>}, {gamble, ギャンブル<*GYANBURU*>, 賭け<*KAKE*>}) and

({競馬<*KEIBA*>, race}, {ギャンブル<*GYANBURU*>, gamble})

are also unified into

({race, レース<*REESU*>, 競争<*KYOUSOU*>, 競馬<*KEIBA*>}, {gamble, ギャンブル<*GYANBURU*>, 賭け<*KAKE*>}).

All word sense associations do not have a counterpart. One possible reason is the disparity in topical coverage between the corpora of the two languages; the other is that

the sense-vs.-clue correlation calculation module produces conflicting sense-vs.-clue correlation matrices. In the former case, word sense associations not having a counterpart should be maintained as is. In the latter case, word sense associations not having a counterpart should be rejected. That is, word sense association ($S(x_1, i_1)$, $S(x_2, i_2)$) is rejected if there is word sense association ($S(y_1, j_1)$, $S(y_2, j_2)$) such that

$$x_1 \in \sum_j S(y_1, j),\ y_1 \in \sum_i S(x_1, i),\ x_2 \in \sum_j S(y_2, j), \text{ and } y_2 \in \sum_i S(x_2, i),$$

but there is no word sense association ($S(y_1, j_1)$, $S(y_2, j_2)$) such that

$$x_1 \in S(y_1, j_1),\ y_1 \in S(x_1, i_1),\ x_2 \in S(y_2, j_2), \text{ and } y_2 \in S(x_2, i_2).$$

[Example]

Assume that an English word association (race, gamble) results in word sense association

({race, 人種<*JINSHU*>, 民族<*MINZOKU*>}, {gamble, ギャンブル<*GYANBURU*>, 賭け<*KAKE*>}),

while a Japanese word association (レース<*REESU*>, ギャンブル<*GYANBURU*>) results in word sense association

({レース<*REESU*>, race}, {ギャンブル<*GYANBURU*>, gamble}).

In this case, both word sense associations are rejected because they conflict.

2.3 Discussion

2.3.1 Potential impact of word sense association networks on natural language processing tasks

Word sense association networks could be applied to several tasks.

(1) Machine translation

WSD is essential for translation-word selection, one of the most fundamental issues in machine translation. WSD based on a word sense association network that determines the senses of words in the first-language input text would restrict candidate translations. From the restricted number of candidates, the most appropriate translations can be selected based on co-occurrence statistics in the second language.

(2) Information retrieval

A word sense association network used as an interface would enable users to input unambiguous queries. That is, users could express their interest unambiguously to the system by indicating the appropriate word senses presented by the system. Such an interface would be effective for both monolingual and cross-language information retrieval because the system would deal with word senses defined as sets of bilingual synonyms.

In addition, a word sense association network would enable queries to be expanded effectively. Query expansion based on word associations often reduces the precision of retrieval because not only associated words relevant to the user's interest, but also associated words irrelevant to the user's interest, are added to the original query. In contrast, query expansion based on word sense associations does not reduce the precision of retrieval because only associated words relevant to the user's interest are added.

(3) Text classification

A word sense association network would enable documents to be characterized using word sense vectors, not word vectors. The similarity between documents based on word sense vectors is likely to be more reliable than that based on word vectors. Therefore, classification accuracy can be improved. It would also enable bilingual text classification since the word sense vectors are common to both languages. Moreover, documents in one language could be classified using a classifier trained on documents in another language.

(4) Question answering

A word sense association network would enable questions to be paraphrased, thereby improving the probability that passages relevant to a question are successfully retrieved. In addition, cross-language question answering would be enabled by paraphrasing across languages.

(5) Text summarization

A word sense association network could be used for multilingual, multi-document summarization because it absorbs lexical differences between related documents.

2.3.2 Linkage of word sense association network with WordNet

The word sense association network and taxonomy-type semantic lexicons like WordNet are complementary. The former provides topical relations between word senses while the latter provide hypernymy/hyponymy relations. Since NLP tasks often require both types of relations, linking a word sense association network with a taxonomy-type semantic lexicon would be beneficial. Topical relations are more suitably acquired from corpora automatically than collected manually, while hypernymy/hyponymy relations are difficult to acquire from corpora. Therefore, linking a word sense association network with a hand-made taxonomy-type semantic lexicon is a worthy undertaking, but one that goes beyond this thesis.

It is obvious that the word sense association network is compatible with WordNet (Miller 1990; Fellbaum 1998), as shown in Fig. 2.5. The word sense association

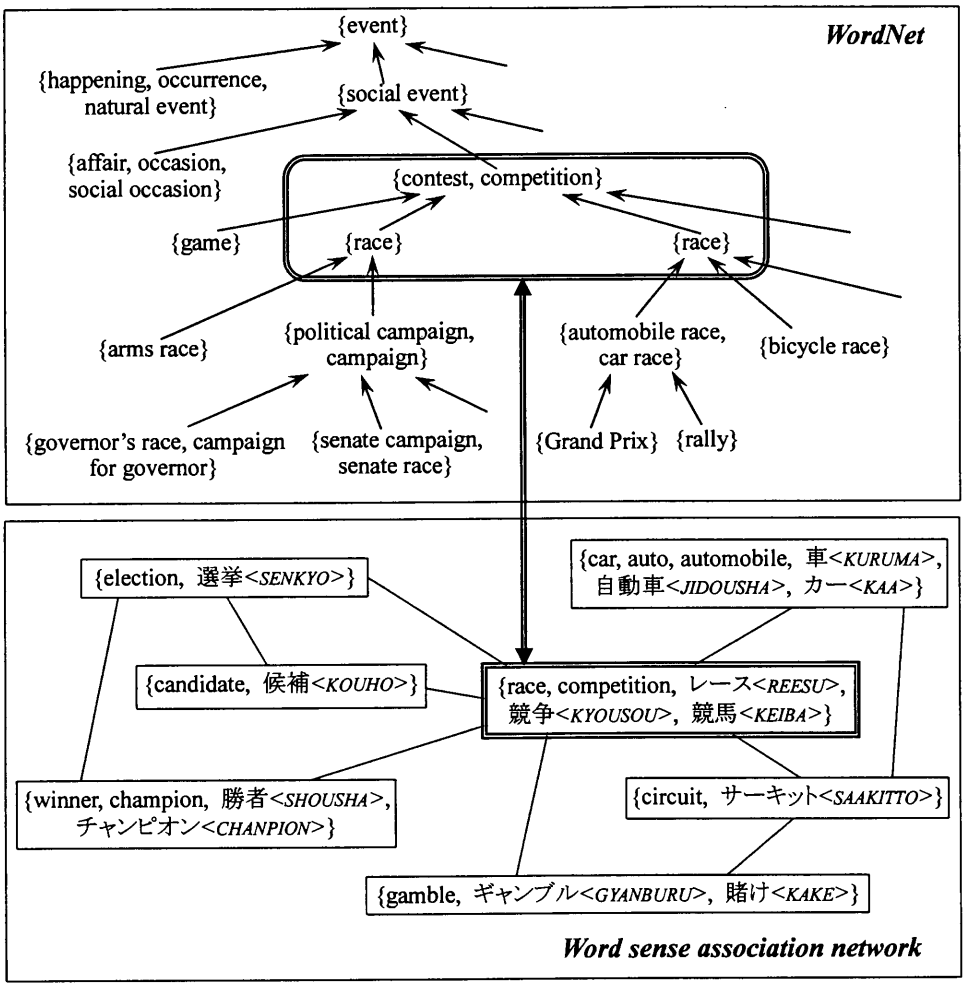


Figure 2.5 Linkage of word sense association network with WordNet

network defines word senses using sets of bilingual synonyms, which are extension of WordNet synsets. Of course, the granularity of word senses often differs between the word sense association network and WordNet. Moreover, WordNet can contain two or more synsets consisting of the same words but representing different concepts (e.g., {race} in Fig. 2.5). Therefore, while a rather sophisticated linkage procedure is needed, it is possible to link word senses in a word sense association network with those in WordNet semi-automatically. This linking process also enables discovery of word senses that are not presently in WordNet. It should be added that the word sense association network is incompatible with other semantic lexicons, such as the EDR concept dictionary and MindNet, in which word senses are defined using descriptive text.

Finally, we mention related work aimed at augmenting WordNet with topical

relations. Harabagiu, et al. (1999) proposed a method for analyzing glosses of synsets to extract topical relations between synsets. That is, the words in the gloss of each synset are disambiguated, and the corresponding synsets are linked with that synset. Agirre, et al. (2001) proposed a method for extracting a set of related words for a synset, called "topic signature," from the World Wide Web. That is, a query is constructed for each synset, using its gloss to retrieve relevant documents from the Web, and the words appearing most distinctively in the retrieved documents are collected. Both these methods and ours acquire topical relations between word senses. The difference is that they do not acquire word senses themselves, while our method acquires word senses as well as topical relations between them.

2.4 Summary

We described a word sense association network. Its nodes represent senses, each of which is defined using a set of bilingual synonyms, and its edges represent associations or topical relations. We then proposed a framework for producing the network from a bilingual comparable corpus and a bilingual dictionary together with an intermediate form of the network that consists of sense inventories and sense-vs.-clue correlation matrices. The basic idea is to align word association networks in respective languages produced from the respective language part of the corpus. It is implemented by interleaving the sense-vs.-clue correlation calculation and the translation equivalent clustering.

We mentioned the potential impact of the word sense association network on various NLP tasks. We also mentioned the possibility of linking the network and WordNet, which are complementary.

Chapter 3

Extraction of Translation Equivalents Based on Contextual Similarity

3.1 Goal and approach

The performance of the proposed method for producing word sense association networks depends on the coverage of the bilingual dictionary used to align the word associations. Methods enabling bilingual dictionaries to be augmented at a reasonable cost are thus required. Our goal was to develop a method for extracting pairs of translation equivalents from bilingual corpora automatically.

There has been a great deal of work on extracting word translations from bilingual corpora automatically. The methods developed so far have advantages and disadvantages. The statistical methods (for example, Gale and Church 1991a; Dagan, et al. 1993) can extract various kinds of word translations. In addition, most of the statistical methods do not require any resource other than a bilingual corpus. However, they usually require a very large corpus, and most of them presuppose that the corpus has been aligned sentence for sentence. The linguistic methods (for example, Yamamoto and Sakamoto 1993; Ishimoto and Nagao 1994) do not require a large corpus and are applicable to an unaligned corpus. However, they are unable to extract pairs of simple words, and they need a bilingual dictionary of simple words.

In response to this situation, we propose a new method having the following characteristics. First, it can extract various kinds of pairs, including mixed pairs of simple and compound words. Second, it is applicable to unaligned bilingual corpora. Given that real-world bilingual corpora usually contain not only many-to-many but also one-to-null sentence correspondences, applicability to unaligned corpora is very important. Third, it can deal with rather small bilingual documents separately.

The proposed method is outlined in Fig. 3.1. The essence of the method is to assess the similarity between words of two languages based on the sets of words co-occurring with them. While it needs a bilingual dictionary of basic words, like the linguistic methods, this does not diminish its usefulness given the wide availability of electronic bilingual dictionaries. The basic idea and the algorithm will be described in detail in

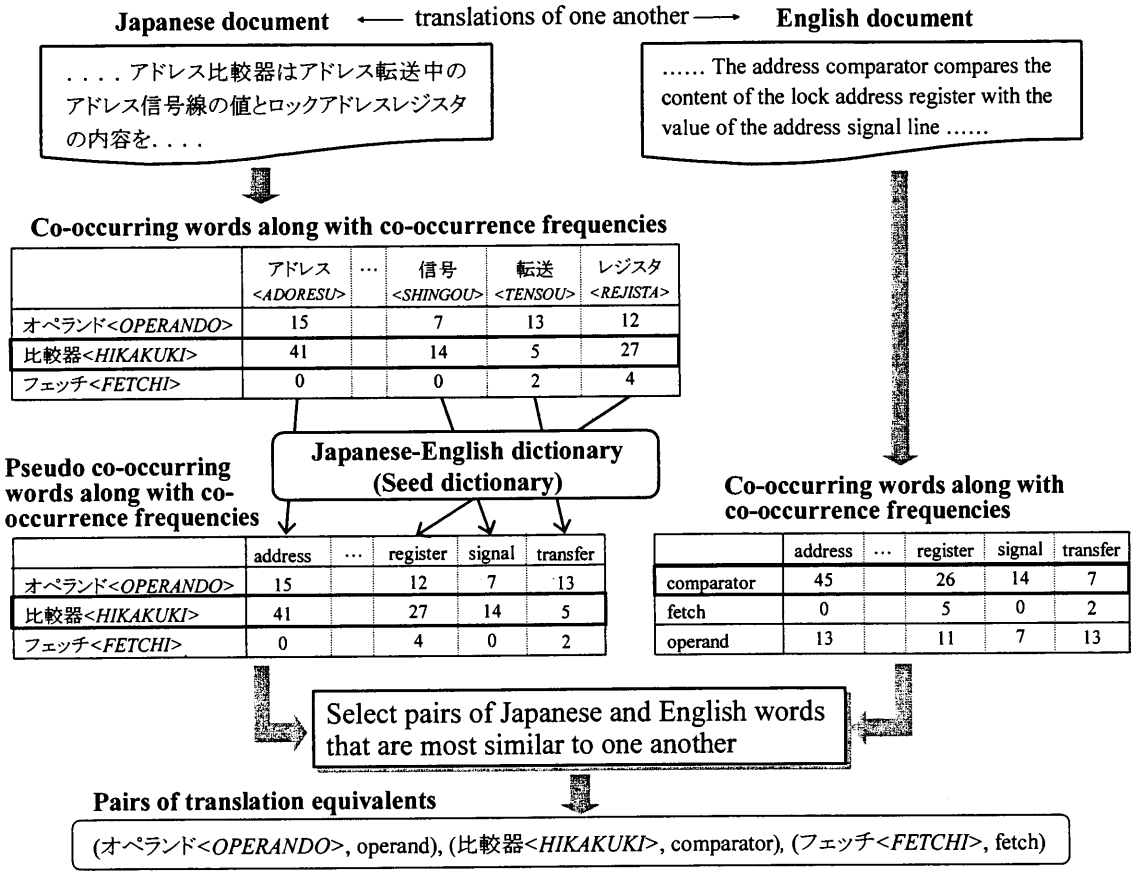


Figure 3.1 Outline of proposed method for extracting translation equivalents

Sections 3.2 and 3.3, respectively.

3.2 Basic idea

3.2.1 Similarity of contexts in two languages

A bilingual corpus consists of pairs of texts in one language and another language, both of which describe the same contents. Therefore, words that correspond to each other are characterized by the same context. If we could evaluate the similarity between these words using the similarity of the contexts characterizing them, we should find that each pair of corresponding words has high similarity. However, the contexts characterizing the words are represented in different languages. To overcome this difficulty, we use an existing bilingual dictionary as a seed dictionary; we evaluate the similarity of the contexts by identifying a word in one language and a word in the other language that are possible translations of each other. Although the identification is not always correct, the overall evaluation of the similarity is sufficiently reliable.

Another problem is that neighboring words are characterized by the same context. The context of a word is always similar not only to that of its counterpart but also to those of counterparts of words occurring in its neighborhood. To avoid this confusion, we treat all instances of a word collectively. That is, we accumulate the contexts of all instances for each word. A word that occurs in the neighborhood of one instance usually does not occur in the neighborhood of all instances. Therefore, by accumulating the contexts of all instances, we can distinguish the words from each other.

The context characterizing a word is usually represented by the set of words co-occurring with it. Therefore, we represent the accumulated context using the set of co-occurring words along with their co-occurrence frequencies, which is more informative than a union of the sets of co-occurring words characterizing the instances. Furthermore, we use the Jaccard coefficient as the similarity measure. That is, similarity $\alpha(x, y)$ between Japanese word x and English word y is defined as

$$\alpha(x,y) = \frac{|CO(x) \cap CO(y)|}{|CO(x) \cup CO(y)|}$$
$$= \frac{|CO(x) \cap CO(y)|}{|CO(x)| + |CO(y)| - |CO(x) \cap CO(y)|},$$

where $CO(x)$ denotes the set of co-occurring words along with their co-occurrence frequencies that characterizes x. Note that the set operations between $CO(x)$ and $CO(y)$ involve coupling of a Japanese word in $CO(x)$ with an English word in $CO(y)$ using a bilingual dictionary, which differs from ordinary set operations. The reliability of the similarity ultimately depends on the coverage of the seed bilingual dictionary. We presuppose that a bilingual dictionary containing at least several thousand basic words is available.

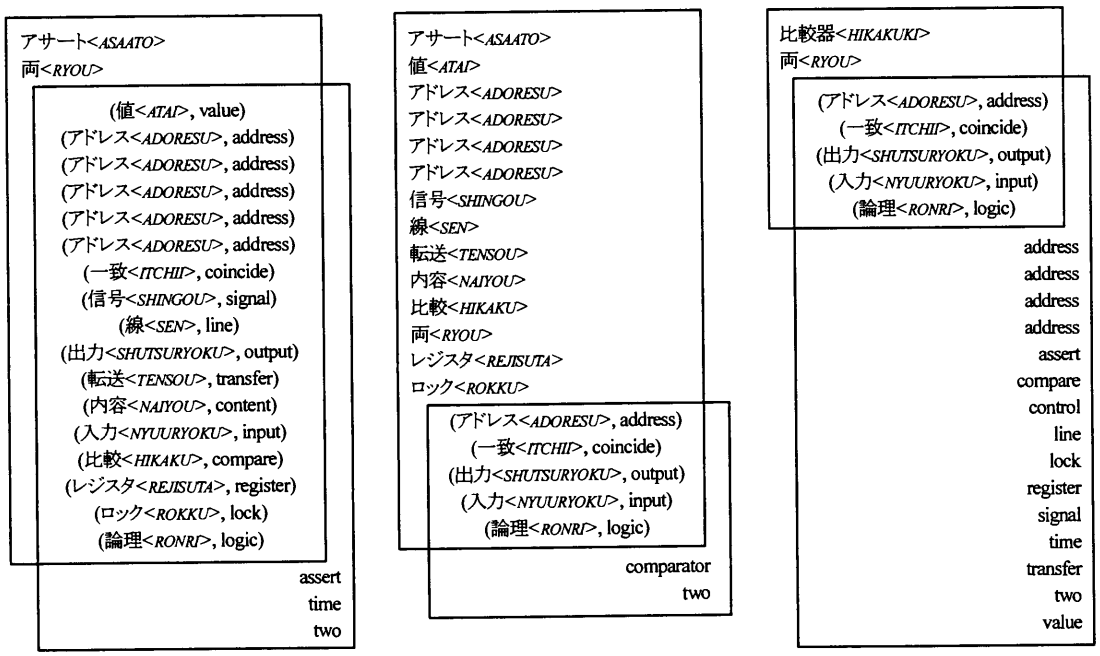
We should mention the class of words with which we deal. That is, the similarity is calculated only between content words, i.e., nouns, verbs, adjectives, and adverbs. Members of the set of co-occurring words along with their co-occurrence frequencies are also restricted to content words. These are because function words, i.e., prepositions/postpositions, auxiliary verbs, and conjunctions, simply play grammatical roles—they do not represent context.

We next give an example of calculating similarity α. Figure 3.2(a) is a small example bilingual corpus. We assume that two pairs of translation equivalents, (比較器 *<HIKAKUKI>*, comparator) and (アサート*<ASAATO>*, assert), are missing in the bilingual dictionary. Figure 3.2(b-1) shows the two sets of co-occurring words along with their co-occurrence frequencies that characterize "比較器*<HIKAKUKI>*" and "comparator,"

アドレス比較器はアドレス転送中のアドレス信号線の値とロックアドレスレジスタの内容を比較する．アドレス比較器の両入力が一致するとその出力には論理“1”がアサートされる．

The address comparator compares the content of the lock address register with the value of the address signal line at the time of address transfer. When the two inputs to the address comparator coincide, logic “1” is asserted as the output.

(a) Example bilingual corpus



α(比較器<*HIKAKUKI*>, comparator)
$= \frac{17}{19+20-17} = 0.77$

(b-1) “比較器<*HIKAKUKI*>” vs. “comparator”

α(比較器<*HIKAKUKI*>, assert)
$= \frac{5}{19+7-5} = 0.24$

(b-2) “比較器<*HIKAKUKI*>” vs. “assert”

α(アサート<*ASAATO*>, comparator)
$= \frac{5}{7+20-5} = 0.23$

(b-3) “アサート<*ASAATO*>” vs. “comparator”

[Note 1] Words occurring in the same sentence are regarded as co-occurring.
[Note 2] Coupling of Japanese word *a* and English word *b* using a bilingual dictionary is denoted as (*a*, *b*).

Figure 3.2 Similarity between sets of co-occurring words along with co-occurrence frequencies

and the similarity between them. Likewise, Fig. 3.2(b-2) shows the similarity between “比較器<*HIKAKUKI*>” and “assert,” and Fig. 3.2(b-3) shows the similarity between “アサート<*ASAATO*>” and “comparator.” We see that the pair of “比較器<*HIKAKUKI*>” and “comparator” has high similarity compared to the other pairs. This suggests that “比較器<*HIKAKUKI*>” and “comparator” are translation equivalents of one another.

3.2.2 Types of co-occurrence

A crucial issue in our method is which definition of "co-occurrence" to use. The selection criterion is how well the contexts assigned to words corresponding to each other overlap. We have several alternatives, including co-occurrence in a sentence, co-occurrence in a window, and syntactic co-occurrence. There are advantages and disadvantages to each.

(1) Co-occurrence in a sentence

Words that occur in the same sentence as a particular word are regarded as co-occurring words of that word. In the example shown in Fig. 3.3(a), the particular word is italicized, and the co-occurring content words are underlined. If we use co-occurrence in a sentence, one-to-one sentence correspondence always results in completely overlapping contexts being assigned to words corresponding to each other. However, one-to-many or many-to-many sentence correspondence always results in partially overlapping contexts being assigned to words corresponding to each other.

(2) Co-occurrence in a window

Words that occur within a certain distance of a particular word are regarded as co-occurring words of that word, as illustrated in Fig. 3.3(b). Co-occurrence in a window possibly avoids the problem of one-to-many and many-to-many sentence correspondence. However, it is difficult to determine the appropriate window size. Although a relatively small window would make it easier to distinguish words from each other, it would often result in contexts with less overlap being assigned to words corresponding to each other. This problem is serious in the case of a pair of languages with different word orders, like Japanese and English.

(3) Syntactic co-occurrence

Words that have a syntactic relation with a particular word are regarded as co-occurring words of that word, as illustrated in Fig. 3.3(c). Syntactic dependency is a strict relation; it enables words to be effectively distinguished from each other. However, use of syntactic co-occurrence would reduce the robustness of the method, as syntactic dependency is not always parallel between languages, especially those with different sentence structures, like Japanese and English.

We use co-occurrence in a sentence because the language-pair we use is Japanese-English. The probability of partially overlapping contexts being assigned to words corresponding to each other is not high, since a larger part of a bilingual corpus usually maintains one-to-one sentence correspondence; moreover, the contexts of all instances of a word are accumulated.

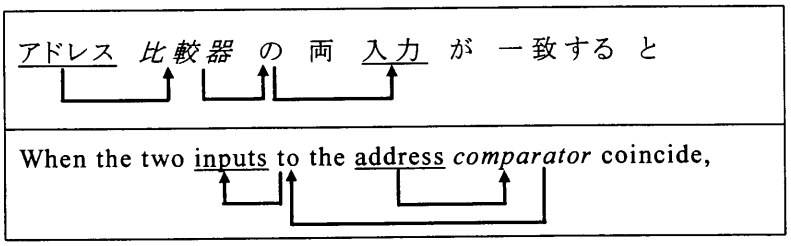
アドレス 比較器 は アドレス 転送 中 の アドレス 信号 線 の 値と ロック アドレス レジスタ の 内容 を 比較する. アドレス *比較器* の 両 入力 が 一致する と その 出力 に は 論理 "1" が アサートさ れる.
The address comparator compares the content of the lock address register with the value of the address signal line at the time of address transfer. When the two inputs to the address *comparator* coincide, logic "1" is asserted as the output.

(a) Co-occurrence in a sentence

アドレス 比較器 は アドレス 転送 中 の アドレス 信号 線 の 値と ロック アドレス レジスタ の 内容 を 比較する. アドレス *比較器* の 両 入力 が 一致する と その 出力 に は 論理 "1" が アサートさ れる.
The address comparator compares the content of the lock address register with the value of the address signal line at the time of address transfer. When the two inputs to the address *comparator* coincide, logic "1" is asserted as the output.

[Note] Window size is set to seven words, excluding function words.

(b) Co-occurrence in a window

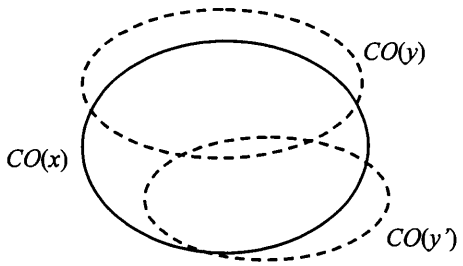


(c) Syntactic co-occurrence

Figure 3.3 Types of co-occurrence

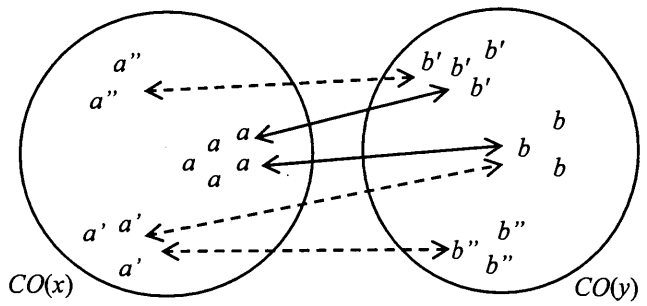
3.2.3 Handling pairs of corresponding documents separately

Pairs of translation equivalents for the same word conflict with each other and are difficult to extract. Let the bilingual corpus contain two pairs of translation equivalents, (x, y) and (x, y'). Then, the set of co-occurring words along with their co-occurrence frequencies for y corresponds to a part of that for x, and the set of co-occurring words along with their co-occurrence frequencies for y' corresponds to another part of that for x as illustrated in Fig. 3.4(a). As a result, neither $\alpha(x, y)$ nor $\alpha(x, y')$ becomes very high,



"Neither α(x, y) nor α(x, y') is very high."

(a) Conflicting pairs of translation equivalents in a corpus



"It is difficult to determine the optimum coupling of the co-occurring words."

(b) Conflicting pairs of translation equivalents contained in bilingual dictionary

Figure 3.4 Conflict between pairs of translation equivalents

and neither (x, y) nor (x, y') is likely to be extracted.

The conflicting pairs of translation equivalents also make it difficult to precisely calculate the similarity between words of two languages. Let the bilingual dictionary contain two pairs of translation equivalents, (a, b) and (a, b'). Further, let the set of co-occurring words along with their co-occurrence frequencies for x include a, and let the set of co-occurring words along with their co-occurrence frequencies for y include both b and b'. To calculate the similarity between x and y, the method described in Subsection 3.2.1 must determine how many instances of a should be coupled with instances of b and how many should be coupled with instances of b'. This cannot be done locally because the conflicts usually form a chain, as illustrated in Fig. 3.4(b).

To avoid the difficulties caused by conflicts among pairs of translation equivalents, the pairs of corresponding bilingual documents are handled separately. This is based on the "one translation per document" hypothesis; i.e., a word has only one sense in a document, and all instances of it are translated the same. It is obvious that conflicts do not occur when the hypothesis is satisfied, and this hypothesis is often true, at least for

technical terms. Note that this strategy contradicts the convention in corpus-based NLP, i.e., the larger the corpus, the higher its effectiveness.

Assume that "コンデンサ<*KONDENSA*>" is translated into "capacitor" and "condenser" in bilingual documents A and B, respectively. When documents A and B are handled separately, it is likely that pairs of equivalent words (コンデンサ<*KONDENSA*>, capacitor) and (コンデンサ<*KONDENSA*>, condenser) are extracted, respectively from A and B. When A and B are handled together, however, it is likely that neither (コンデンサ<*KONDENSA*>, capacitor) nor (コンデンサ<*KONDENSA*>, condenser) is extracted, or at best either (コンデンサ<*KONDENSA*>, capacitor) or (コンデンサ<*KONDENSA*>, condenser) is extracted.

Handling pairs of corresponding documents separately is advantageous from the operational viewpoint. It enables bilingual dictionaries to be continually enhanced by adding new entries extracted from newly received documents. In relation to this, our view on performance should be mentioned. We attach more importance to precision than to recall because, even if a pair of a word and its translation cannot be extracted from a document, it can likely be extracted from a following document. In addition, handling pairs of corresponding documents separately enables us to neglect computational efficiency—the computer program only has to process a small corpus in a reasonable amount of time.

3.3 Algorithm

3.3.1 Outline

As shown in Fig. 3.5, our proposed method consists of three steps.

1) Extract Japanese co-occurrences

First, the Japanese text is segmented into sentences, the sentences are divided into words, the pairs of words co-occurring in each sentence are extracted, and their co-occurrence frequencies are counted. A set of co-occurring words along with their co-occurrence frequencies is thereby produced for each Japanese word.

2) Extract English co-occurrences

Step 1) is applied to the English text, and a set of co-occurring words along with their co-occurrence frequencies is produced for each English word.

3) Extract pairs of translation equivalents

The similarity between the Japanese and English words is calculated pair-wisely, and the pairs of words with the mutually highest similarity are selected. This step is performed again after the extracted pairs of words are fed back.

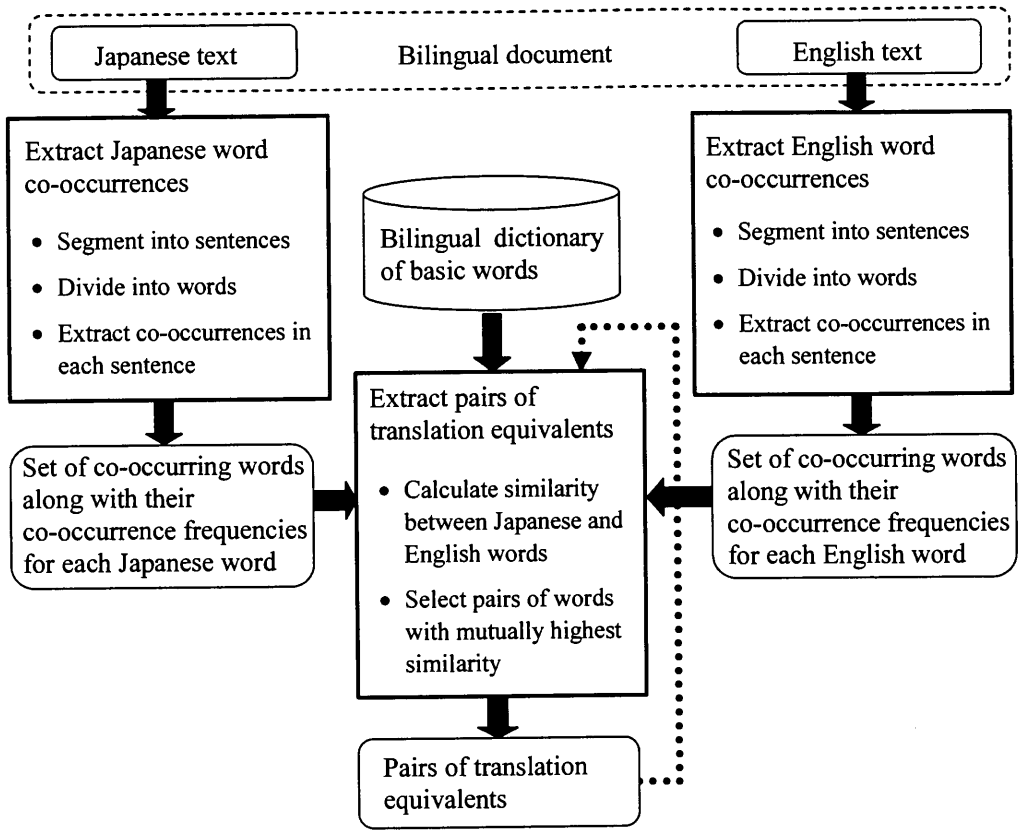


Figure 3.5 Flow of translation equivalent extraction

3.3.2 Extraction of words

All content words, i.e., nouns, verbs, adjectives, adverbs, and unknown words, most of which are probably nouns, are extracted. Compound nouns are also extracted. A compound noun is defined by the following part-of-speech sequence patterns, where N, ADJ, UK, and NP denote noun, adjective, unknown word, and compound noun, respectively, and '+' means the preceding symbol can be repeated many times.

- Japanese compound noun

  NP := {N|UK} {N|UK}+

- English compound noun

  NP := {N|UK|ADJ} {N|UK}+

Note that only maximal NPs are extracted; non-maximal NPs are rejected. An NP is maximal when it is not included in a larger NP; otherwise, it is called a non-maximal NP. The non-maximal NPs are rejected because there is uncertainty about whether they are compound nouns. Exceptionally, an English NP starting with N is extracted even when

it is included in a NP starting with ADJ because the ADJ may be a modifier.

In the following subsections, extracted Japanese type-words are denoted as x_1, x_2, $\cdots$, x_m, and extracted English type-words are denoted as y_1, y_2, $\cdots$, y_n. Type-words having a stem with the same spelling but taking different parts of speech are regarded as one type-word. For example, the noun "増加<*ZOUKA*>" and verb "増加する<*ZOUKASURU*>" are merged into one; likewise, the noun "increase" and verb "increase" are merged into one. This is done because part-of-speech disambiguation cannot be done with sufficient accuracy, especially for English. Since our main purpose is to extract pairs of words that are translations of one another, addressing parts of speech is a secondary issue.

3.3.3 Extraction of word co-occurrence

For each word, all words that occur in the same sentence as it are extracted, and the co-occurrence frequencies of the co-occurring words are counted. Although they are included in the same sentence, a compound word and its constituent words should not be regarded as co-occurring. Therefore, the constituent words are excluded from the set of co-occurring words along with their co-occurrence frequencies that characterizes a compound word, and vice versa.

The resultant set of co-occurring words along with their co-occurrence frequencies for the i-th Japanese word, x_i, is denoted as

$$CO(x_i) = \{x_k / f_{i,k} \mid k = 1,2,\cdots,m\} \quad (i = 1,2,\cdots,m),$$

where $f_{i,k}$ is the co-occurrence frequency of x_i and x_k. Likewise, the resultant set of co-occurring words along with their co-occurrence frequencies for the j-th English word, y_j, is denoted as

$$CO(y_j) = \{y_k / g_{j,k} \mid k = 1,2,\cdots,n\} \quad (j = 1,2,\cdots,n),$$

where $g_{j,k}$ is the co-occurrence frequency of y_j and y_k.

3.3.4 Calculation of similarity between words

As mentioned in Subsection 3.2.3, conflicts among pairs of translation equivalents makes it difficult to precisely calculate the similarity between words in two languages. Considering computational efficiency, we calculate the similarity approximately as follows.

i) Modify the set of co-occurring words along with their co-occurrence frequencies

Words having no counterparts are eliminated from every set of co-occurring words along with their co-occurrence frequencies as follows.

-If $\forall y_j \quad (x_i, y_j) \notin D$, then

$$CO(x_k) \leftarrow CO(x_k) - \{x_i / f_{ki}\} \qquad (i=1, 2, \cdots, m; k=1, 2, \cdots, m).$$

-If $\forall x_i \quad (x_i, y_j) \notin D$, then

$$CO(y_k) \leftarrow CO(y_k) - \{y_j / g_{kj}\} \qquad (j=1, 2, \cdots, n; k=1, 2, \cdots, n).$$

The D denotes the seed bilingual dictionary, a collection of possible pairs of translation equivalents. Words having no counterparts reduce the values of the similarity irregularly depending on their co-occurrence frequencies. Eliminating them makes the similarity more reliable.

ii) Convert the Japanese set of co-occurring words along with their co-occurrence frequencies into a pseudo set of co-occurring words along with their co-occurrence frequencies

Co-occurrences with Japanese words are translated into pseudo co-occurrences with English words. The frequency of pseudo co-occurrences with an English word is the sum of the frequencies of co-occurrences with Japanese words that can be translated into the English word. That is,

$$CO'(x_i) = \{y_j / f'_{i,j} \mid j = 1,2,\cdots,n\},$$

where $f'_{i,j} = \sum_{(x_k, y_j) \in D} f_{i,k} \qquad (i=1, 2, \cdots, m).$

iii) Calculate similarity of Japanese words and English words

The similarity of x_i and y_j is calculated as follows:

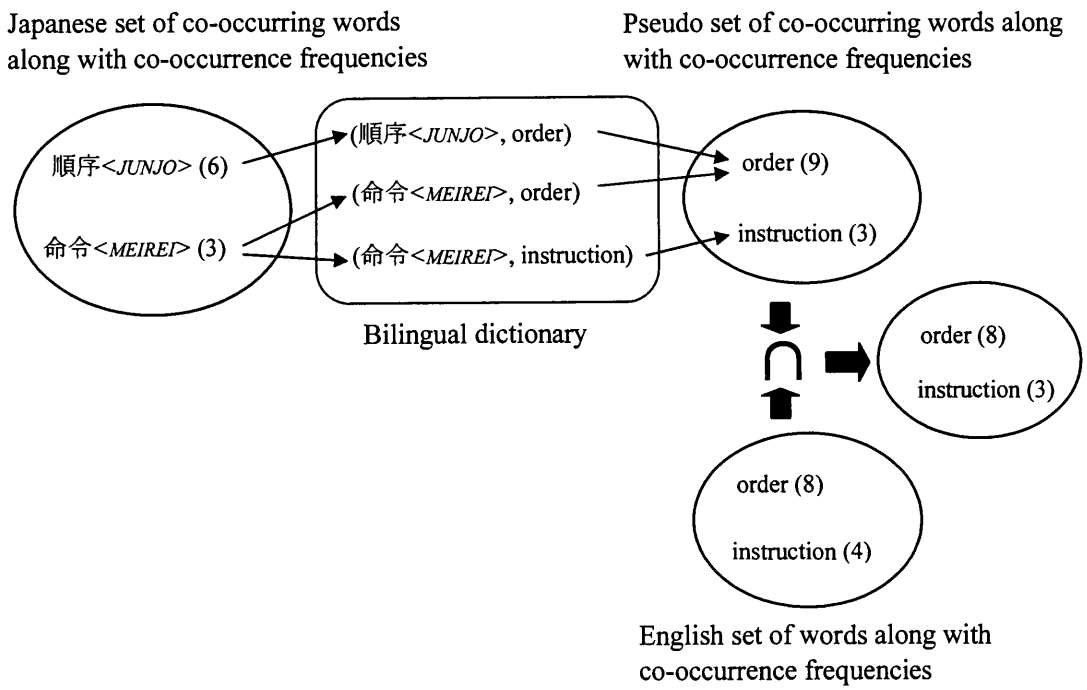
$$\alpha(x_i, y_j) = \frac{|CO(x_i) \cap CO(y_j)|}{|CO(x_i)| + |CO(y_j)| - |CO(x_i) \cap CO(y_j)|}$$

$$\cong \frac{|CO'(x_i) \cap CO(y_j)|}{|CO(x_i)| + |CO(y_j)| - |CO'(x_i) \cap CO(y_j)|}$$

$$= \frac{\sum_k \min\{f'_{i,k}, g_{j,k}\}}{\sum_k f_{i,k} + \sum_k g_{j,k} - \sum_k \min\{f'_{i,k}, g_{j,k}\}}.$$

$$(i=1, 2, \cdots, m; j=1, 2, \cdots, n)$$

This formula approximates the similarity defined in Subsection 3.2.1; i.e., the



[Note] Figures in parentheses after words are frequencies.

Figure 3.6 Example of overestimated intersection of sets of co-occurring words

intersection of the set of co-occurring words along with their co-occurrence frequencies for x_i and the set of co-occurring words along with their co-occurrence frequencies for y_j is replaced with the intersection of the "pseudo" set of co-occurring words along with their co-occurrence frequencies for x_i and the set of co-occurring words along with their co-occurrence frequencies for y_j.

The conversion in step ii) often generates excessive pseudo co-occurrences due to the conflicts among pairs of translation equivalents. Let (x_p, y_q) and (x_p, y_r) be conflicting pairs of translation equivalents. Then, $f_{i,p}$ co-occurrences of x_i with x_p are converted into $f_{i,p}$ pseudo co-occurrences of x_i and y_q and into $f_{i,p}$ pseudo co-occurrences of x_i and y_r, resulting in a total of $2 \cdot f_{i,p}$ co-occurrences. As a result, the intersection of the Japanese set of co-occurring words along with their co-occurrence frequencies and the English set of co-occurring words along with their co-occurrence frequencies is overestimated.

Figure 3.6 shows an example, where (命令<*MEIREI*>, order) and (命令<*MEIREI*>, instruction) are conflicting pairs of translation equivalents. The three "命令<*MEIREI*>"s

in the Japanese set of co-occurring words along with their co-occurrence frequencies are translated into three "order"s and three "instruction"s in the pseudo set of co-occurring words along with their co-occurrence frequencies. The six "順序<*JUNJO*>"s in the Japanese set of co-occurring words along with their co-occurrence frequencies are converted into six "order"s. As a result, the intersection of the Japanese set of co-occurring words along with their co-occurrence frequencies and the English set of co-occurring words along with their co-occurrence frequencies includes eight "order"s and three "instruction"s, which is an overestimation.

The overestimation could be avoided if the conversion were done using a translation probability matrix whose (i, j)-element shows the probability that x_i is translated into y_j. However, we do not use one because it is difficult to estimate the translation probabilities. A translation probability matrix unmatched to the target bilingual documents would underestimate the intersection of the Japanese set of co-occurring words along with their co-occurrence frequencies and the English set of co-occurring words along with their co-occurrence frequencies.

Finally, we mention that the approximation error caused by excessively generated pseudo co-occurrences is not serious. Again, let (x_p, y_q) and (x_p, y_r) be conflicting pairs of translation equivalents. For simplicity, assume that neither (x_p, y_q) nor (x_p, y_r) conflicts with any other pair of translation equivalents. In this case, an error may occur when calculating the similarity of x_i co-occurring with x_p and y_j co-occurring with both y_q and y_r. The contribution of (x_p, y_q) and (x_p, y_r) to $\sum_k \min\{f'_{i,k}, g_{j,k}\}$ in the formula in step iii) is $\min\{f_{i,p}, g_{j,q}\} + \min\{f_{i,p}, g_{j,r}\}$, while its correct value is $\min\{f_{i,p}, g_{j,q}+g_{j,r}\}$. It is overestimated when $f_{i,p} < g_{j,q}+g_{j,r}$. However, even if it is overestimated, it is less than $g_{j,q}+g_{j,r}$. When (x_i, y_j) is a correct pair of translation equivalents, $f_{i,p}$ is probably nearly equal to $g_{j,q}+g_{j,r}$, so the contribution of (x_p, y_q) and (x_p, y_r) to $\sum_k \min\{f'_{i,k}, g_{j,k}\}$ is probably nearly equal to $g_{j,q}+g_{j,r}$. As a result, the estimated similarity is probably the maximum for the correct pair of translation equivalents.

In short, excessively generated pseudo co-occurrences do not necessarily result in overestimation of the similarity. Furthermore, overestimation does not necessarily reverse the order of the similarity values.

3.3.5 Selection of pairs with mutually highest similarity

Every pair of Japanese word x_i and English word y_j is selected as a pair of translation equivalents when it meets the following conditions:

(a) $\forall k(\neq j) \quad \alpha(x_i, y_j) > \alpha(x_i, y_k)$ and $\forall k(\neq i) \quad \alpha(x_i, y_j) > \alpha(x_k, y_j)$

(b) $\forall k(\neq j) \quad (x_i, y_k) \notin D$ or $\alpha(x_i, y_k) = 0$ and $\forall k(\neq i) \quad (x_k, y_j) \notin D$ or $\alpha(x_k, y_j) = 0$

(c) $\alpha(x_i, y_j) \notin D$.

Condition (a) means that pairs with the mutually highest similarity are selected. Two or more translation equivalents are never extracted for one word. This corresponds to the "one translation per document" hypothesis. Condition (b) means pairs conflicting with those contained in the seed bilingual dictionary are excluded. This condition is particularly severe, but we prefer precision to recall, as mentioned in Subsection 3.2.3. Condition (c) means that pairs already in the bilingual dictionary are excluded.

3.3.6 Feedback of extracted pairs of translation equivalents

The performance of our method depends on how well the seed bilingual dictionary covers the corpus, as mentioned in Subsection 3.2.1. Generally, the wider the coverage, the more reliable the similarity. Accordingly, extracted pairs of translation equivalents are fed back to the bilingual dictionary, and the procedures described in Subsections 3.3.4 and 3.3.5 are carried out again. The pairs that are fed back, of course, include erroneous ones, which may degrade performance. The experiments described in the following section include evaluation of the effect of feedback.

3.4 Experimental evaluation

3.4.1 Method and materials

The proposed method was evaluated experimentally using patent specification documents written in Japanese and English. Five pairs of corresponding documents in the semi-conductor field were separately used, following the strategy described in Subsection 3.2.3. The dictionary for a Japanese-English machine translation system was used as the seed bilingual dictionary. It contains approximately 60,000 Japanese entry words with an average of 3.8 English translation equivalents per word.

Evaluation was done by comparing the pairs of translation equivalents extracted from each pair of documents with those extracted manually. Two sets of recall and precision, one before feedback and the other after feedback, were calculated. Recall is the proportion of pairs of translation equivalents included in the corpus that our method successfully extracted, i.e.,

$$R_{TR} = \frac{|TE_A \cap TE_M|}{|TE_M|}.$$

Precision is the proportion of extracted pairs of translation equivalents that were actually correct, i.e.,

$$P_{TR} = \frac{|TE_A \cap TE_M|}{|TE_A|}.$$

The TE_A denotes the set of pairs of translation equivalents extracted by our method, and TE_M denotes the set of manually extracted pairs of translation equivalents that had not been entered into the seed bilingual dictionary.

The manual extraction of pairs of translation equivalents was done on an instance basis. The criterion for identifying compound nouns, described in Subsection 3.3.2, was also applied to manual extraction—maximal NPs were extracted and non-maximal NPs were rejected. However, when a non-maximal compound noun was preferable to the maximal one, the maximal one was replaced with the non-maximal one. For example, "回路素子数<*KAIRO-SOSHI-SUU*>" ("the number of circuit elements") was replaced with "回路素子<*KAIRO-SOSHI*>" ("circuit element"). In addition, compound nouns whose part-of-speech sequence pattern is more complex than those described in Subsection 3.3.2 were also extracted. For example, "carry look ahead circuit" (part-of-speech sequence pattern "Noun+Verb+Adverb+Noun") was extracted.

3.4.2 Characteristics of patent specification documents

The English patent specification documents had been produced by translating the Japanese patent specification documents. However, the English documents had been enhanced and sometimes supplemented by the inventors and/or patent attorneys, while the Japanese counterpart had been left untouched. As a result, the pairs of Japanese and English documents were hard to align sentence by sentence.

A quantitative profile of the five pairs of documents is shown in Table 3.1. Characteristic values C1 through C9, which were calculated based on the results of manual extraction of pairs of translation equivalents, summarize the word correspondence between corresponding documents.

- Value C4, the percentage of pairs of translation equivalents that were difficult to extract due to a conflict with other pairs, was 17.4% on average.

Table 3.1 Profile of patent specification documents used for evaluation

		Document	(i)	(ii)	(iii)	(iv)	(v)	Total[†)]
A1	Japanese text	No. of sentences	90	120	686	230	178	1,304
A2		No. of content words	1,322	2,089	8,023	3,846	2,449	17,729
A3		Avg. sentence length (A2/A1)	14.7	17.4	11.7	16.7	13.8	13.6
A4		No. of distinct content words	202	273	719	392	524	2,110
A5		Avg. word frequency (A2/A4)	6.5	7.7	11.2	9.8	4.7	8.4
B1	English text	No. of sentences	94	143	704	236	178	1,355
B2		No. of content words	1,463	2,055	9,561	4,326	2,872	20,277
B3		Avg. sentence length (B2/B1)	15.6	14.4	13.6	18.3	16.1	15.0
B4		No. of distinct content words	244	312	936	485	629	2,606
B5		Avg. word frequency (B2/B4)	6.0	6.6	10.2	8.9	4.6	7.8
C1	Correspondences between words	No. of distinct pairs of translation equivalents	211	316	1,008	608	660	2,803
C2		No. of unknown pairs	75	126	417	315	302	1,235
C3		No. of unknown pairs that conflict with others	12	19	85	45	54	215
C4		Ratio of C3 to C2 (%)	16.0	15.1	20.4	14.3	17.9	17.4
C5		No. of pairs registered in dictionary	136	190	591	293	358	1,568
C6		No. of pairs registered in dictionary that share the Japanese word with other pairs	24	28	193	65	75	385
C7		Ratio of C6 to C5 (%)	17.6	14.7	32.7	22.2	20.9	24.6
C6'		No. of pairs registered in dictionary that share the English word with other pairs	25	40	245	79	74	463
C7'		Ratio of C6' to C5 (%)	18.4	21.1	41.5	27.0	20.7	29.5
C8		No. of pairs of simple words	163	221	714	343	438	1,879
C9		Coverage of dictionary, C5/C8 (%))	83.4	86.0	82.8	85.4	81.7	83.4

[†)] Values (i) through (v) were simply averaged without considering inter-document overlap of words or pairs of translation equivalents.

- Values C7 and C7' show the ratios of pairs of translation equivalents that caused overgeneration of pseudo co-occurrences. Value C7, which was relevant when the

Japanese sets of co-occurring words along with their co-occurrence frequencies were converted into pseudo sets of co-occurring words along with their co-occurrence frequencies, as described in Subsection 3.3.4, was 24.6% on average. In contrast, C7', which was relevant when the English sets of co-occurring words along with their co-occurrence frequencies were converted into pseudo sets of co-occurring words along with their co-occurrence frequencies, was 29.5% on average. The first case, which resulted in less overgeneration, was used in the experiment.

- Value C9, the coverage of the bilingual dictionary over the test documents, was 83.4% on average.

It should be added that we neglected reference numbers in the documents. The underlined numbers in the following pair of sentences are examples of reference numbers.

...アドレス比較器504の両入力が一致すると....

...the two inputs to address comparator 504 coincide with....

Achieving correspondence between reference numbers, although trivial, would improve the performance of our method. However, since reference numbers are specific to patent documents, we neglected them to generalize the evaluation.

3.4.3 Results

The experimental results are summarized in Table 3.2. Averaged over the five pairs of documents, recall was 30.5% and precision was 74.7% before feedback, and 33.8% and 76.7% after feedback. An additional experiment in which feedback was repeated one more time showed that doing so resulted in no further improvement. Given that the pairs of translation equivalents to be extracted included those with a frequency equal to one, 33.8% recall and 76.7% precision are reasonable. Our method is thus effective in reducing the cost of bilingual dictionary augmentation.

The low recall is compensated for by our strategy to handle rather small bilingual documents separately. To prove the validity of this strategy, we counted the number of pairs of translation equivalents that were not extracted from one of the five documents but were extracted from the remaining four documents. If we used this approach, recall would be improved 2.7% (see rows F1 and F2 of Table 3.2). Recall could thus be substantially improved by processing many documents serially.

Example extracted pairs of translation equivalents are shown below, with unknown simple words underlined. We see that our method can extract various types of pairs of translation equivalents.

Table 3.2 Recall and precision of translation equivalent extraction

	Document		(i)	(ii)	(iii)	(iv)	(v)	Total
C2	No. of unknown pairs of translation equivalents in document		75	126	417	315	302	1,235
D1	Before feedback	No. of extracted pairs	31	53	190	131	100	505
D2		No. of correct extracted pairs	22	46	144	96	69	377
D3		Recall, D2/C2 (%)	29.3	36.5	34.5	30.5	22.8	30.5
D4		Precision, D2/D1 (%)	71.0	86.8	75.8	73.3	69.0	74.7
E1	After feedback	No. of extracted pairs	31	60	202	140	111	544
E2		No. of correct extracted pairs	23	50	157	102	85	417
E3		Recall, E2/C2 (%)	30.7	39.7	37.6	32.4	28.1	33.8
E4		Precision, E2/E1 (%)	74.2	83.3	77.7	72.9	76.6	76.7
F1	No. of pairs recovered using other four documents		6	6	9	4	8	33
F2	Substantial improvement in recall, F1/C2 (%)		8.0	4.8	2.2	1.3	2.6	2.7

- Pairs of simple words
  (排気<*HAIKI*>, pump)
  (引き続き<*HIKITSUZUKI*>, subsequently)
  (フェッチ<*FETCHI*>, fetch)
  (容量<*YOURYOU*>, capacitance)
- Pairs of compound words
  (ガス供給機構<*GASU-KYOUKYUU-KIKOU*>, gas supplier)
  (桁上げ生成回路<*KETAAGE-SEISEI-KAIRO*>, carry generation circuit)
  (高周波加熱<*KOUSHUUHA-KANETSU*>, radio frequency heating)
- Mixed pairs of a Japanese simple word and an English compound word
  (圧損<*ASSON*>, pressure loss)
  (液面<*EKIMEN*>, liquid level)
  (薄膜<*HAKUMAKU*>, thin film)
- Mixed pairs of a Japanese compound word and an English simple word
  (気化器<*KIKA-KI*>, vaporizer)
  (接続口<*SETSUZOKU-GUCHI*>, connector)
  (熱処理<*NETSU-SHORI*>, anneal)

The advantage of handling documents separately is shown by the extraction of the translation equivalent pair (容量<*YOURYOU*>, capacitance). If the bilingual documents had not been handled separately, this pair would not have been extracted due to its conflict with the dominant translation equivalent pair, (容量<*YOURYOU*>, capacity).

3.5 Discussion

3.5.1 Advantages of proposed method

Our method uses both co-occurrence frequencies and a bilingual dictionary of basic words, making it a hybrid approach. However, it differs from both statistical and linguistic methods. The novelty and resulting advantages of our method are discussed below.

In conventional statistical methods, words are characterized by their occurrence frequencies or occurrence positions in a corpus. In our method, words are characterized by their context or the sets of words co-occurring with them. The sets of co-occurring words, which provide far richer information than the occurrence frequencies or positions, enable pairs of translation equivalents to be extracted from a rather small bilingual corpus. Pairs of translation equivalents that occur only a few times in the corpus can be extracted.

In conventional linguistic methods, a bilingual dictionary of basic words is used to examine the constituent-level correspondence between each pair of words. In our method, a bilingual dictionary is used to evaluate the contextual similarity between words in the two languages, resulting in a difference in the types of pairs of translation equivalents. That is, while conventional linguistic methods can extract only pairs of compound words, our method can extract pairs of simple words, pairs of compound words, and even mixed pairs of simple and compound words.

3.5.2 Performance compared with previous methods

It is difficult to compare the performances of different methods because of the different task settings. Some methods extract pairs of compound words while others extract pairs of simple words. Methods using a bilingual dictionary extract unknown pairs of translation equivalents while methods without a bilingual dictionary extract all pairs of translation equivalents. Different corpora of different language-pairs are used for evaluation. We chose two previous methods that had been evaluated under relatively similar settings.

Kumano and Hirakawa (1994) proposed a method that uses a bilingual dictionary of basic words and the occurrence frequencies of words. They applied it to a

Japanese-English corpus of patent specification documents to extract pairs of compound words and pairs of unknown words. The precision of extracting pairs of translation equivalents weighted by the frequencies of Japanese words was 72.9% for 3,224 compound nouns and 54.0% for 389 unknown words.

To compare our results with those of Kumano and Hirakawa, we classified the pairs of translation equivalents extracted in the experiment described in Section 3.4 into three groups:

(a') pairs of a Japanese compound noun and its English equivalent,

(b') pairs of a Japanese unknown simple word and its English equivalent, and

(c') pairs of a Japanese known simple word and its English equivalent.

Then, we counted the frequencies of Japanese words and calculated the precision weighted by the frequencies for each group. The results were as follows:

(a') 88.7% for 1,737 compound nouns,

(b') 90.6% for 414 unknown simple words, and

(c') 91.4% for 209 known simple words.

We compared these results with those of Kumano and Hirakawa, taking the difference in corpus size (1,304 vs. 2,128 sentences, respectively) into account. Although the ratio of Japanese words whose English equivalents were determined by our method was slightly less than that with their method, our precision was higher than theirs. In particular, our method was much better for simple words.

Fung (1995) proposed extracting pairs of translation equivalents from an unaligned bilingual corpus. She applied her purely statistical method based on the positions of words in a corpus to an English-Chinese corpus to extract English common nouns and proper nouns and then determine their Chinese equivalents. The precision was 73.1% for 661 English words whose frequency was two or more. The method could not determine Chinese equivalents for most of the 2,118 English words whose frequency was one.

To compare Fung's method with our method, we estimated recall using her method for English words whose frequency was two or more. Under the assumption that each English word has one and only one Chinese equivalent in the corpus, recall is equal to precision, that is, 73.1%. In addition, we classified pairs of translation equivalents extracted in the experiment described in Section 3.4 into two groups: (i) pairs of a Japanese word with a frequency equal to or larger than two and its English equivalent and (ii) pairs of a Japanese word with a frequency equal to one and its English equivalent. Then, we calculated the recall and precision for each group. Recall was 32.7% and precision was 82.5% for Japanese words with a frequency equal to or larger

than two and 35.2% and 70.5% for Japanese words with a frequency equal to one. Comparing the results for the first group with Fung's results showed that our method had lower recall and higher precision. The results for the second group showed that our method is effective even for words with a frequency equal to one.

3.5.3 Limitations and directions for extension

While our method is effective, as demonstrated experimentally, it has several problems we need to address in future work. Some of them are discussed below with possible extensions.

(1) Refining the compound-noun extraction procedure

The simplified procedure described in Subsection 3.3.2 often omits a compound noun as well as extracts an inappropriate string of words. One reason is that it neglects non-maximal compound nouns. For example, it failed to extract compound nouns "回路素子<*KAIRO-SOSHI*>" ("circuit element") and "絶縁耐圧<*ZETSUEN-TAIATU*>" ("dielectric strength"), and instead extracted maximal compound nouns "回路素子数<*KAIRO-SOSHI-SUU*>" ("number of circuit elements") and "絶縁耐圧向上<*ZETSUEN-TAIATU-KOUJOU*> ("improved dielectric strength"). Another reason, especially for English compound-noun extraction, is that too simple part-of-speech sequence patterns are used. For example, it could not extract compound nouns such as "carry look ahead circuit" and "air-operated valve."

To extract non-maximal compound nouns precisely, global processing, e.g., use of N-gram frequencies, is required. As for extracting English compound nouns, shallow parsing seems to be required. Some incorrect pairs of equivalent words extracted in the experiment (e.g., (回路素子数<*KAIRO-SOSHI-SUU*>, circuit element)) were partially correct. This implies that refining the compound-noun extraction procedure would considerably improve recall and precision.

(2) Combining with other methods

To show that contextual similarity is useful for extracting pairs of translation equivalents from unaligned bilingual corpora, we tested its use without combining it with other methods. Obviously, however, combining it with other methods will improve recall and precision. In particular, it is natural to combine our method with conventional linguistic methods that extract pairs of compound words by evaluating their constituent-level correspondence.

3.6 Related work

Rapp (1995) was the first to discuss the possibility of extracting translation equivalents

based on similarity of co-occurrence patterns. His original idea was to permute the rows, and synchronously columns, of a word association matrix of one language to maximize its similarity to that of the other language. He showed that matrix similarity depends on the ratio of corresponding rows representing correct pairs of translation equivalents. However, he found that matrix permutation is computationally inefficient, and no further results have been reported.

Tanaka and Iwasaki (1996) used a similar idea to choose the best translation equivalent from a small set of candidates corpus-dependently. The essence of their method is to construct the translation probability matrix that minimizes the distance between the word association matrix of the first language and that of the second language. It cannot extract a pair of translation equivalents not in the bilingual dictionary.

Although it was developed independently, our method described can be considered a practical implementation of Rapp's idea. Using a bilingual dictionary of basic words as seed pairs of translation equivalents, it reduces a large number of permutations to a much smaller number of comparisons between sets of co-occurring words. We demonstrated that the method is useful for extracting translation equivalents from document-aligned but not sentence-aligned corpora (Kaji and Aizono 1996).

Methods for extracting translation equivalents from comparable or unrelated corpora were subsequently proposed (Fung and McKeown 1997; Fung and Yee 1998; Rapp 1999). All of them use the same framework as ours along with modification to cope with difference in size and topics between language texts. Table 3.3 summarizes the characteristics of these methods together with the evaluation results. The results revealed the limitation of using the co-occurrence pattern, although it is very useful for extracting translation equivalents from comparable corpora. Further work is needed to enable the extraction of translation equivalents from comparable corpora. We believe that our choice of document-aligned corpora as source corpora is appropriate from the practical point of view; there are many bilingual documents that are difficult to align sentence by sentence, e.g., patent documents, paper abstracts, product manuals, and Web pages.

Using co-occurrence patterns is not the only way to extract translation equivalents from comparable corpora. Nakagawa (2001) demonstrated that extracting compound noun translations based on the correspondence between their constituent words can be applied to comparable corpora; the essence of his method is to disambiguate compound noun translations extracted from comparable corpora by evaluating their termhood. While this is much more effective than using the similarity of co-occurrence patterns for

Table 3.3 Methods for extracting translation equivalents from comparable corpora

		Fung and McKeown (1997)	Fung and Yee (1998)	Rapp (1999)	cf. Our method (Kaji and Aizono 1996)
Co-occurrence		Co-occurrence in a paragraph	Co-occurrence in a sentence	Co-occurrence in a small window	Co-occurrence in a sentence
Values of elements of co-occurring word vector		Weighted mutual information score	Term frequency – inverse document frequency	Log-likelihood ratio	Co-occurrence frequency
Similarity measure		Cosine measure	Cosine measure	City-block metric	Weighted Jaccard coefficient
Seed pairs of translation equivalents		Restricted to mid-frequency words	Weighted according to order of translations	Restricted to pairs of a word and its first translation	All
Evaluation	Corpus	Japanese newspaper *Nihon Keizai Shimbun* (127 MB) and English newspaper *Wall Street Journal* (49 MB)	Chinese newspaper *Mingpao* (8.8 MB) and English newspaper *Hong Kong Standard* (3 MB)	German newspaper *Frankfurter Allgemine Zeitung* (135 million words) and English newspaper *Guardian* (163 million words)	Six pairs of Japanese and English patent documents that are translations of one another
	Number of test words	19	–	100 [2)]	1,235 [3)]
	Recall	–	–	–	33.8%
	Precision	About 30% [1)]	–	72%	76.7%

1) Poor results even though candidate translation equivalents were limited to a few hundred words.

2) Test words were common words, and almost all correct translations were already in seed bilingual dictionary. Therefore, results are meaningless from practical point of view.

3) Evaluation was done on all pairs of translation equivalents found in corpus but not in seed bilingual dictionary.

compound noun translations in which there is correspondence between the constituent words, it is useless for simple noun translations as well as for compound noun translations in which there is no correspondence between the constituent words.

3.7 Summary

We developed a method for extracting pairs of translation equivalents from a bilingual corpus based on contextual similarity. First, for each word in both languages, the set of co-occurring words, along with their co-occurrence frequencies, is extracted from the

corpus. Then, similarities between sets of co-occurring words are calculated pair-wisely using a bilingual dictionary of basic words. Finally, pairs of words with the mutually highest similarity are selected.

The effectiveness of this method was demonstrated experimentally using Japanese and English patent specification documents; recall was 33.8% and precision was 76.7%. The method, which uses both co-occurrence information given by a corpus and bilingual knowledge given by an existing dictionary of basic words, has advantages specific to statistical and to linguistic methods. First, it can extract a variety of word translations, including pairs of simple words, pairs of compound words, and mixed pairs of simple and compound words. Second, it is applicable to bilingual corpora that are difficult to align sentence by sentence. Third, it does not require a very large corpus, so it enables rather small bilingual corpora to be handled separately.

Chapter 4

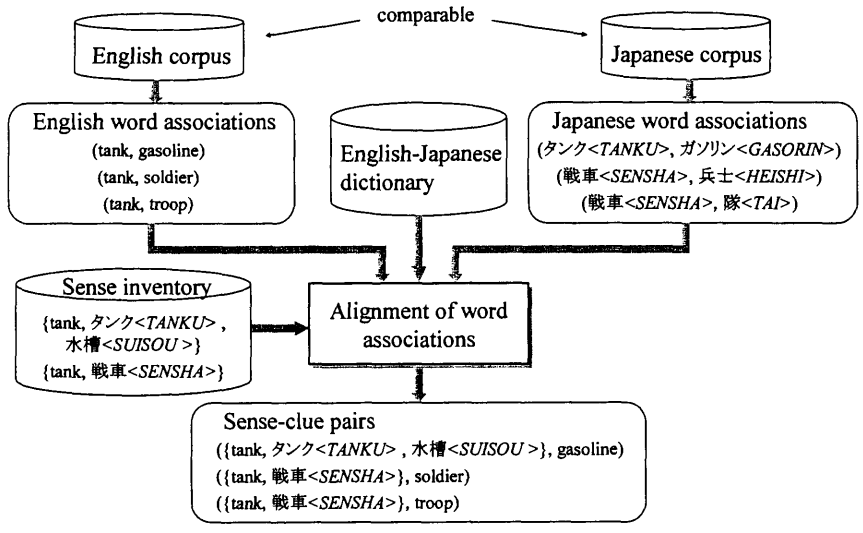
Iterative Calculation of Sense-vs.-Clue Correlations Based on Translingual Alignment of Word Associations

4.1 Goal and approach

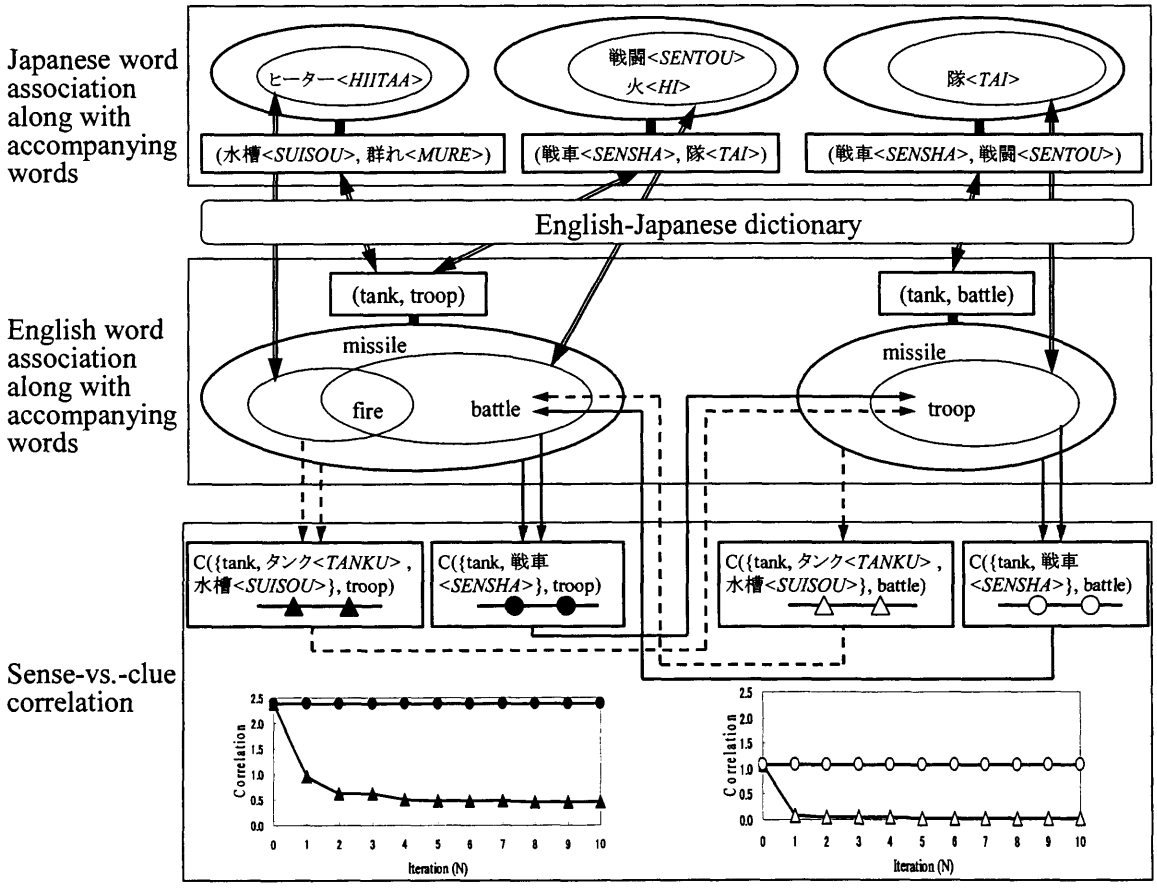
Our goal was to develop a method for calculating correlations between the senses of a polysemous word and the clues identifying the sense of that word. The task of calculating sense-vs.-clue correlations is considered to be a learning stage for WSD, and, in general, various techniques for WSD can be applied. However, our overall objective imposes a restriction on the approach—a fully unsupervised learning method must be used. That is, we also want to automate word sense acquisition. The sense-vs.-clue correlation calculation method is used with automatically acquired and defined senses. Therefore, not only must the training corpus not be sense-tagged manually but supplementary information must not be used to bootstrap the learning process, e.g., textual definitions of senses cannot be used as seeds (Yarowsky 1995; Karov and Edelman 1998).[3]

WSD techniques using bilingual corpora meet the requirement for fully unsupervised learning. However, those using parallel corpora have a critical deficiency—the availability of large parallel corpora is extremely limited. Using a second-language monolingual corpus and a bilingual dictionary instead of a parallel corpus solves this problem (Dagan and Itai 1994). Although we used this approach, Dagan and Itai's original method has a number of shortcomings. First, it does not have a learning stage; it only disambiguates instances of first-language polysemous words by using the statistical information of the second-language corpus. We needed to develop a mechanism to map the second-language word associations onto the first-language sense-vs.-clue correlations inversely. Second, it is hampered by the sparseness of co-occurrence data as well as the uncertain correspondence between the two language texts. Novel ideas were required to overcome these difficulties.

[3] Whether a method is fully unsupervised depends on how word senses are defined. Methods using textual definitions of senses as seeds are not fully unsupervised when word senses are defined as sets of translation equivalents. They are of course fully unsupervised with word senses defined using descriptive texts.



(a) Alignment of word associations



(b) Alignment of word associations along with their accompanying words and iterative calculation of sense-vs.-clue correlations

Figure 4.1 Outline of method proposed for calculating correlations between senses and clues

Our proposed method is outlined in Fig. 4.1. The basic idea is to align word associations by using a bilingual dictionary and convert each alignment into a pair of a sense and a clue identifying the sense (Fig. 4.1(a)). However, this naive method suffers from ambiguous alignment. In addition, it fails to align many of the word associations with their counterparts due to the disparity in topical coverage between corpora of the two languages. To overcome these problems, we assume that the correlation between a sense and a clue depends on those between that sense and related clues. That is, each word association is characterized by a set of accompanying words, i.e., words associated with both words making up the association, and the correlation between a sense and a clue is calculated iteratively using the correlations between that sense and the accompanying words that are also clues (Fig. 4.1(b)). The basic idea and the algorithm will be described in detail in Sections 4.2 and 4.3, respectively.

We add that we restrict both the target words and clues to nouns. In other words, only associations between nouns are dealt with. Associations between nouns and verbs /adjectives are not dealt with, although they are also significant. Noun-noun associations, which are topical ones, can simply be extracted from a corpus based on the frequency of their co-occurrence in a window. On the other hand, extraction of noun-verb/adjective associations, which are syntactic ones, requires parsing of sentences. Taking the less availability of robust parsers into account, we focus on applying our method to noun-noun associations. Its application to noun-verb and noun-adjective associations is left as future work.

4.2 Basic idea

4.2.1 Translingual alignment of word associations

Unlike with a parallel corpus, we can align neither sentences nor words between first- and second-language texts making up a comparable corpus. However, we can assume that translations of words that are associated in one language are also associated in the other language. Based on this assumption, we extract a collection of word associations from each language text independently of the other language text, and then translingually align the word associations using a bilingual dictionary.

Aligning word associations enables us to acquire pairs of a sense and a clue identifying it. That is, the alignment of first-language word association (x, x') with second-language word association (y, y') suggests that x' is a clue identifying the sense of target word x translated into y. Therefore, the alignment of (x, x') with (y, y') can be converted into sense-clue pair ($\{x, y\}$, x'), where $\{x, y\}$ denotes the sense of x that can be translated into y. For example, the alignment of (tank, gasoline) with (タンク<*TANKU*>,

ガソリン<*GASORIN*>) suggests that "gasoline" is a clue identifying the "container" sense of "tank," which can be translated as "タンク<*TANKU*>,"[4] and the alignment of (tank, soldier) with (戦車<*SENSHA*>, 兵士<*HEISHI*>) suggests that "soldier" is a clue identifying the "military vehicle" sense of "tank," which can be translated as "戦車<*SENSHA*>."

In this framework of translingually aligning word associations, we encounter two major problems: alignment ambiguity and alignment failure due to disparity in topical coverage between the corpora of the two languages as well as by incomplete coverage of the bilingual dictionary. The following subsections discuss how we overcome these problems.

4.2.2 Coping with alignment ambiguity

Word association matching using a bilingual dictionary often results in a word association in one language being aligned with two or more word associations in the other language. For example, the English word association (tank, troop) is aligned with the Japanese word associations (水槽<*SUISOU*>, 群れ<*MURE*>), (槽<*SOU*>, 多数<*TASUU*>), (戦車<*SENSHA*>, 群<*GUN*>), (戦車<*SENSHA*>, 多数<*TASUU*>), and (戦車<*SENSHA*>, 隊<*TAI*>).[5]

We overcome alignment ambiguity by assuming that word associations making up a correct alignment are accompanied by many words that can be aligned with each other. That is, if the alignment of first-language word association (x, x') with second-language word association (y, y') is correct, most words associated with both x and x' can be aligned with words associated with both y and y'. Based on this assumption, the plausibility of a word association alignment, or the plausibility of a sense given by a clue, is evaluated. Then, the correlation between the sense and the clue is calculated as the product of the correlation between the target word and the clue and the (normalized) plausibility factor. It should be noted that two or more alignments may suggest the same pair of a sense and a clue. In such cases, the maximum plausibility factor of those alignments is taken. Thus, the sense-vs.-clue correlation is defined as follows:

(correlation between sense $\{x, y\}$ and clue x')

= (correlation between x and x') ·

---

[4] While several English-Japanese dictionaries render "タンク<*TANKU*>" as a translation of "tank" representing both the "container" sense and the "military vehicle" sense, "タンク<*TANKU*>" is rarely used to represent the latter sense.

[5] The examples in Section 4.2 are actual examples based on the corpus and bilingual dictionary used in the experiments described in Section 4.5.

(tank, troop)－(水槽<*SUISOU*>, 群れ<*MURE*>)

air, area, fire, government

(tank, troop)－(槽<*SOU*>, 多数<*TASUU*>)

area, army, control, force

(tank, troop)－(戦車<*SENSHA*>, 群<*GUN*>)

area, army, battle, commander, force, government

(tank, troop)－(戦車<*SENSHA*>, 多数<*TASUU*>)

Serb, area, army, battle, force, government

(tank, troop)－(戦車<*SENSHA*>, 隊<*TAI*>)

Russia, Serb, air, area, army, battle, commander, defense, fight, fire, force, government, helicopter, soldier

Figure 4.2 Example sets of translingually alignable accompanying words

$$\max_{y'} \text{(plausibility factor of alignment of } (x, x') \text{ with } (y, y')).$$

[Example]

(Correlation between sense {tank, 戦車<*SENSHA*>} and clue "troop")

= (correlation between "tank" and "troop") ·

max {(plausibility factor of alignment of (tank, troop) with (戦車<*SENSHA*>, 群<*GUN*>)),

(plausibility factor of alignment of (tank, troop) with (戦車<*SENSHA*>, 多数<*TASUU*>)),

(plausibility factor of alignment of (tank, troop) with (戦車<*SENSHA*>, 隊<*TAI*>))}.

Next, we define the plausibility factor of a word association alignment. A naive definition derived from the above assumption is that the factor equals the size of the set of translingually alignable accompanying words that characterizes the alignment, i.e., the number of words accompanying the first-language word association that can be aligned with words accompanying the second-language word association. Figure 4.2 shows sets of translingually alignable accompanying words that characterize the alignments of the English word association (tank, troop) with the Japanese word associations (水槽<*SUISOU*>, 群れ<*MURE*>), (槽<*SOU*>, 多数<*TASUU*>), (戦車<*SENSHA*>, 群<*GUN*>), (戦車<*SENSHA*>, 多数<*TASUU*>), and (戦車<*SENSHA*>, 隊<*TAI*>).

The sizes of these sets are 4, 4, 6, 6, and 14. The set characterizing the correct alignment, i.e., (tank, troop) - (戦車<*SENSHA*>, 隊<*TAI*>), is the largest.

In the above naive definition, all accompanying words are treated equally. However, the set of translingually alignable accompanying words often contains erroneous words. For example, the set characterizing the alignment of (tank, troop) with (槽<*SOU*>, 多数<*TASUU*>) contains "force" (See Fig. 4.2). This is because "force" is associated with both "tank" and "troop," "効果<*KOUKA*>" is associated with both "槽<*SOU*>" and "多数<*TASUU*>," and "force" and "効果<*KOUKA*>" are translations of each other. Actually, "force" associated with both "tank" and "troop" has a "military" sense, which differs from the sense translated into "効果<*KOUKA*>" ("effect"). Therefore, "force" in the set of translingually alignable accompanying words that characterizes the alignment of (tank, troop) with (槽<*SOU*>, 多数<*TASUU*>) is erroneous.

To minimize the effect of erroneous words, we define the plausibility factor of a word association alignment as the sum of the correlations between the sense suggested by the alignment and the translingually alignable accompanying words. Note that the accompanying words are also clues identifying the sense of the target word.

[Example] (See Fig. 4.2)

(Plausibility factor of alignment of (tank, troop) with (槽<*SOU*>, 多数<*TASUU*>))
= (correlation between sense {tank, 槽<*SOU*>} and clue "area")
+ (correlation between sense {tank, 槽<*SOU*>} and clue "army")
+ (correlation between sense {tank, 槽<*SOU*>} and clue "control")
+ (correlation between sense {tank, 槽<*SOU*>} and clue "force").

(Plausibility factor of alignment of (tank, troop) with (戦車<*SENSHA*>, 隊<*TAI*>))
= (correlation between sense {tank, 戦車<*SENSHA*>} and clue "Russia")
+ (correlation between sense {tank, 戦車<*SENSHA*>} and clue "Serb")
+ (correlation between sense {tank, 戦車<*SENSHA*>} and clue "air")

.
.
.

+ (correlation between sense {tank, 戦車<*SENSHA*>} and clue "soldier").

Since the correlations between a sense and erroneous accompanying words tend to be low, the plausibility factor of a word association alignment is evaluated reliably.

Thus, sense-vs.-clue correlations are recursively defined. That is, the sense-vs.-clue correlations are defined based on the plausibility factors of word association alignments, which are, in turn, defined using the sense-vs.-clue correlations. We will describe an iterative algorithm for calculating the sense-vs.-clue correlations in Subsection 4.3.5.

Table 4.1 Examples of English word associations and their Japanese counterparts along with mutual information values

English word association		Japanese word association [mutual information]	
(tank, army)		(槽<*SOU*>, 多数<*TASUU*>[1])	[3.13]
	√	(戦車<*SENSHA*>, 軍<*GUN*>)	[6.78]
		(戦車<*SENSHA*>, 多数<*TASUU*>[1])	[4.89]
	√	(戦車<*SENSHA*>, 隊<*TAI*>)	[5.83]
	√	(戦車<*SENSHA*>, 兵隊<*HEITAI*>)	[7.00]
	√	(戦車<*SENSHA*>, 陸軍<*RIKUGUN*>)	[7.59]
(tank, troop)		(水槽<*SUISOU*>, 群れ<*MURE*>[2])	[6.95]
		(槽<*SOU*>, 多数<*TASUU*>[1])	[3.13]
		(戦車<*SENSHA*>, 群<*GUN*>[3])	[4.53]
		(戦車<*SENSHA*>, 多数<*TASUU*>[1])	[4.89]
	√	(戦車<*SENSHA*>, 隊<*TAI*>)	[5.83]

[Note 1] √: correct counterpart

[Note 2] 1) "a large number," 2) "a group of people or animals," 3) "group"

We next describe an additional idea that can be combined with the above idea. That is, we assume that alignments with strong word associations are preferable to those with weak word associations. This assumption was the basis for Dagan and Itai's (1994) translated-word selection method using a second-language monolingual corpus. However, it may not be reliable, as exemplified in Table 4.1. On the one hand, for (tank, army), the Japanese counterpart with the highest mutual information value, i.e., (戦車<*SENSHA*>, 陸軍<*RIKUGUN*>), is correct. On the other hand, for (tank, troop), the Japanese counterpart with the highest mutual information value, i.e., (水槽<*SUISOU*>, 群れ<*MURE*>), is incorrect. Therefore, we evaluated two alternatives experimentally: one using a plausibility factor based on translingually alignable accompanying words, and the other using a plausibility factor multiplied by the correlation (i.e., mutual information) of the second-language word association.

4.2.3 Coping with alignment failure

In a weakly comparable corpus, the topics covered by the first-language texts do not necessarily coincide with those covered by the second-language texts. Moreover, the bilingual dictionary used for aligning word associations does not cover the complete vocabulary of the corpus. Therefore, a first-language word association is not necessarily aligned with second-language word association(s). Obviously, if the method described

(tank, troop)

Army, Bosnian, Bosnian government, Chechen, Chechnya, Force, Grozny, Israel, Moscow, Mr. Yeltsin, Mr. Yeltsin's, NATO, Pentagon, Republican, Russia, Russian, Secretary, Serb, U.N., Yeltsin, Yeltsin's, air, area, army, assault, battle, bomb, carry, civilian, commander, control, defense, fight, fire, force, government, helicopter, military, missile, rebel, soldier, weapon

Figure 4.3 Example set of accompanying words regardless of translingual alignability

in the preceding subsection is used, a first-language word association does not produce sense-vs.-clue correlations unless it is aligned with second-language counterpart(s). Furthermore, even if a first-language word association is aligned with one or more second-language counterparts, all the alignments may be incorrect. In this case, the clue has zero correlation with the correct sense and non-zero correlations with the incorrect ones.

Disparity in topical coverage between the texts of the two languages and incomplete coverage of the bilingual dictionary also reduce the number of translingually alignable accompanying words even for a correct word association alignment. Obviously, the plausibility factor of a word association alignment based on a small set of translingually alignable accompanying words is not reliable.

To overcome this problem, we evaluate the plausibility of a sense given by a clue in an additional way, considering that a word association and its accompanying words tend to suggest the same sense, whether they can be aligned with counterparts in another language or not. A first-language word association, i.e., a pair of the target word and a clue, is characterized by a set of accompanying words regardless of translingual alignability. This set is shared among all senses of the target word, and, for each sense, the additional plausibility factor is defined as the sum of the correlations between the sense and the accompanying words.

[Example]

English word association (tank, troop) is characterized by the set of accompanying words shown in Fig. 4.3. Using this set, we calculate additional plausibility factors of senses {tank, 槽<*SOU*>} and {tank, 戦車<*SENSHA*>} given by clue "troop" as follows.

(Additional plausibility factor of sense {tank, 槽<*SOU*>} given by clue "troop")
= (correlation between sense {tank, 槽<*SOU*>} and clue "Army")
+ (correlation between sense {tank, 槽<*SOU*>} and clue "Bosnian")
+ (correlation between sense {tank, 槽<*SOU*>} and clue "Bosnian government")

.
.
.

+ (correlation between sense {tank, 槽<*SOU*>} and clue "weapon").

(Additional plausibility factor of sense {tank, 戦車<*SENSHA*>} given by clue "troop")

= (correlation between sense {tank, 戦車<*SENSHA*>} and clue "Army")

+ (correlation between sense {tank, 戦車<*SENSHA*>} and clue "Bosnian")

+ (correlation between sense {tank, 戦車<*SENSHA*>} and clue "Bosnian government")

.
.
.

+ (correlation between sense {tank, 戦車<*SENSHA*>} and clue "weapon").

Most of the accompanying words usually have different correlations with different senses. Therefore, the additional plausibility factors differ among the senses although the set of accompanying words is shared among all the senses. Note that the additional plausibility factors based on sets of accompanying words regardless of translingual alignability are effective only when they are used together with the plausibility factors based on sets of translingually alignable accompanying words. The details of the sense-vs.-clue correlation calculation using both sets of translingual alignable accompanying words and sets of accompanying words regardless of translingual alignability will be described in Subsection 4.3.5.

4.3 Algorithm

4.3.1 Outline

Our proposed sense-vs.-clue correlation calculation method consists of the following steps, as illustrated in Fig. 4.4.

1) Extract a set of word associations from each of the first-language and second-language corpora. Each word association is characterized by a set of accompanying words.
2) Align the first-language word associations with the second-language ones, and characterize each alignment using a set of translingually alignable accompanying words.
3) Calculate the correlations between the senses and clues iteratively based on the sets of translingually alignable accompanying words as well as the sets of

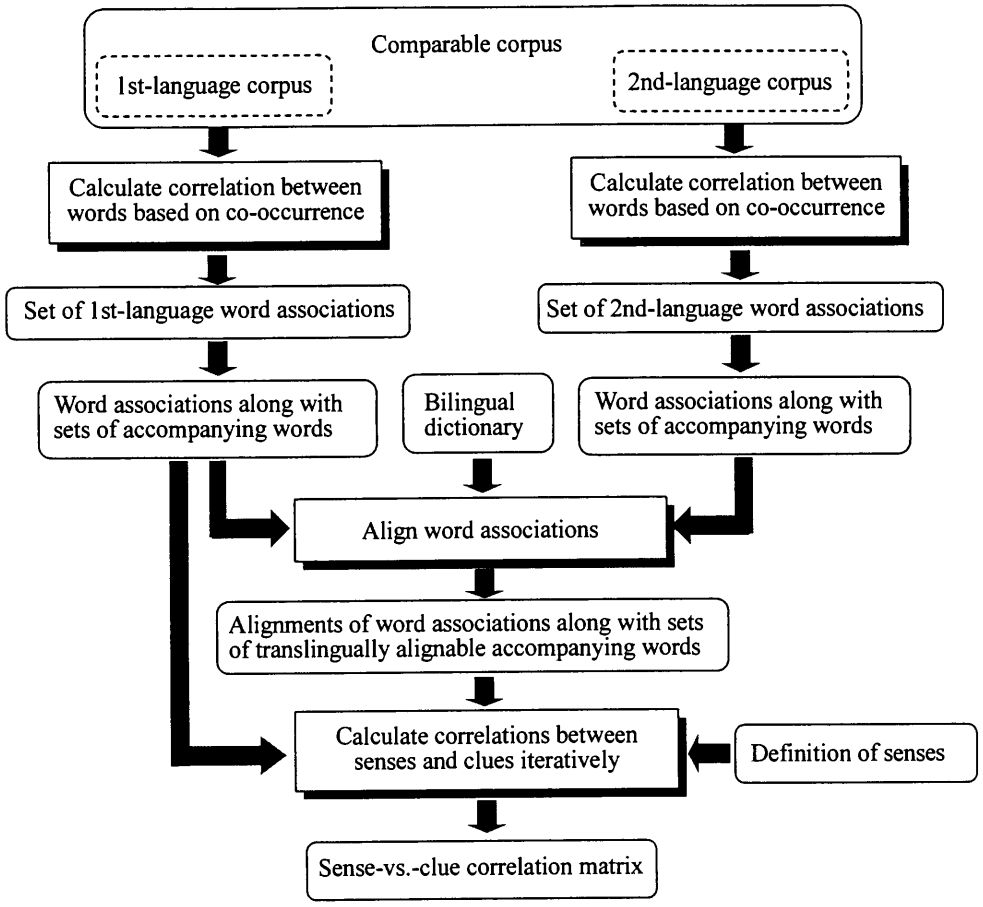


Figure 4.4 Flow of sense-vs.-clue correlation calculation

accompanying words regardless of translingual alignability.

Note that steps 2) and 3) are carried out for each target word, and a sense-vs.-clue correlation matrix is produced for each target word.

4.3.2 Definition of word senses

Our method presupposes that each sense of target word x in the first language is defined as a synonym set consisting of x itself and one or more of its translation equivalents (y_1, y_2, ...) in the second language.[6] The synonym set is similar to the WordNet (Miller 1990) synset except that it is bilingual, not monolingual. Example sets are given below.

{tank, タンク<*TANKU*>, 水槽<*SUISOU*>, 槽<*SOU*>}

[6] While we address the automatic definition of word senses in Chapter 5, here it does not matter whether word senses are defined manually or automatically.

{tank, 戦車<*SENSHA*>}

These synonym sets define the "container" sense and the "military vehicle" sense of "tank," respectively.

Our method is based on the premise that the senses of a polysemous word in a language are lexicalized differently in another language. Therefore, it works best if translation equivalents preserving the ambiguity of the target word are excluded from the synonym sets defining senses. Not only are such translation equivalents useless, but also they can cause confusion. An example is given below.

{title, 肩書き<*KATAGAKI*>, 称号<*SHOUGOU*>, ~~タイトル<*TAITORU*>~~, 敬称<*KEISHOU*>}

{title, 題名<*DAIMEI*>, 題目<*DAIMOKU*>, 表題<*HYOUDAI*>, 書名<*SHOMEI*>, ~~タイトル<*TAITORU*>~~}

{title, ~~タイトル<*TAITORU*>~~, 選手権<*SENSHUKEN*>}

These synonym sets define the "person's rank or profession" sense, the "name of a book or play" sense, and the "championship" sense of "title." The Japanese word "タイトル<*TAITORU*>," which represents all these senses, should be excluded from all these synonym sets.

4.3.3 Extraction of word associations

The corpus of each language is statistically processed in order to extract a collection of word associations in the language (Kaji, et al. 2000). First, words are extracted from the corpus, and their occurrence frequencies are counted. Words with occurrence frequencies less than a predetermined threshold are rejected. Pairs of words co-occurring in a window are also extracted, and their co-occurrence frequencies are counted. In the present implementation, the words are restricted to nouns and unknown words, which are probably nouns, and the window size is set to 25 words, excluding function words.

Next, for each pair of words x and x', mutual information $MI(x, x')$ is calculated:

$$MI(x,x') = \log \frac{Pr(x,x')}{Pr(x) \cdot Pr(x')},$$

where $Pr(x)$ is the occurrence probability of x, and $Pr(x, x')$ is the co-occurrence probability of x and x'. Finally, pairs of words having mutual information larger than a predetermined threshold are selected as word associations. Statistically insignificant pairs are filtered out through a log-likelihood ratio test (Dunning 1993).

It should be added that the mutual information meets the requirement of the succeeding steps. The mutual information values are used as the base for the sense-vs.-clue correlations (Subsection 4.3.5), and the score for each sense of the target

word is defined as the (weighted) sum of the correlations of the sense with clues in the context (Section 4.4). Consider the difference between $MI(t, c_1)$ and $MI(t, c_2)$, where t is the target word, and c_1 and c_2 are clues identifying the sense of t. The difference is represented as

$$MI(t,c_1) - MI(t,c_2) = \log\frac{Pr(t,c_1)}{Pr(t)\cdot Pr(c_1)} - \log\frac{Pr(t,c_2)}{Pr(t)\cdot Pr(c_2)} = \log Pr(t \mid c_1) - \log Pr(t \mid c_2).$$

Therefore, the contributions of clues to determining the sense of the target word depend on the conditional probabilities of the target word given by the clues, whether they occur frequently or not. This seems appropriate because the WSD task is to infer the sense of the target word from clues in the context. Note that, if we used the log-likelihood ratio instead of the mutual information, less-frequent clues would make smaller contributions to determining the sense of the target word.

4.3.4 Translingual alignment of word associations

Let $X(x)$ be the set of clues identifying the sense of first-language target word x. That is,

$$X(x) = \{x' \mid (x, x') \in R_X\},$$

where R_X denotes the collections of word associations extracted from the corpus of the first language. We denote the j-th clue identifying the sense of x as $x'(j)$.

Each first-language word association $(x, x'(j))$ is aligned with all possible second-language word associations. We denote the set consisting of counterparts of $(x, x'(j))$ as $Y(x, x'(j))$. That is,

$$Y(x, x'(j)) = \{(y, y') \mid (y, y') \in R_Y, (x, y) \in D, (x'(j), y') \in D\},$$

where R_Y denotes the collections of word associations extracted from the corpus of the second language, and D denotes a bilingual dictionary, i.e., a collection of pairs consisting of a first-language word and a second-language word that are possible translations of one another.

Then, each first-language word association $(x, x'(j))$ is characterized by a set of accompanying words, denoted as $Z(x, x'(j))$. That is,

$$Z(x, x'(j)) = \{x'' \mid x'' \in X(x), (x'(j), x'') \in R_X\}.$$

Furthermore, each alignment of first-language word association $(x, x'(j))$ with second-language word association (y, y') $(\in Y(x, x'(j)))$ is characterized by a set of translingually alignable accompanying words, denoted as $W((x, x'(j)), (y, y'))$. That is,

$$W((x, x'(j)), (y, y')) = Z(x, x'(j)) \cap \{x'' \mid \exists y''(\in V(y, y'))\ (x'', y'') \in D\},$$

where

$$V(y, y') = \{y'' \mid (y, y'') \in R_Y, (y', y'') \in R_Y\}.$$

4.3.5 Calculation of correlations between senses and clues

Let $S(x, i)$ be the i-th sense of target word x. The correlation between $S(x, i)$ and j-th clue $x'(j)$ is defined as

$$C(S(x,i),x'(j)) = MI(x,x'(j)) \cdot \frac{PL(S(x,i),x'(j))}{\max_k PL(S(x,k),x'(j))},$$

where $MI(x, x'(j))$ is the mutual information of x and $x'(j)$, and $PL(S(x, i), x'(j))$ is the plausibility factor for $S(x, i)$ given by $x'(j)$. The mutual information of the target word and the clue is the base of the correlation between the sense and the clue; it is multiplied by the normalized plausibility factor. Thus, one of the senses has the maximum correlation equal to the mutual information of the target word and the clue.

The plausibility factor is defined as the weighted sum of two component plausibility factors, i.e.,

$$PL(S(x,i),x'(j)) = PL_1(S(x,i),x'(j)) + \alpha \cdot PL_2(S(x,i),x'(j)),$$

where α is a parameter adjusting the relative weights of the component plausibility factors.

The first component plausibility factor, PL_1, is based on the set of accompanying words regardless of translingual alignability. It is defined as the sum of correlations between the sense and the accompanying words, i.e.,

$$PL_1(S(x,i),x'(j)) = \sum_{x'' \in Z(x,x'(j))} C(S(x,i),x'').$$

The second component plausibility factor, PL_2, is based on the set of translingually alignable accompanying words. We have two alternative formulae for defining the second component plausibility factor.

[Formula I]

$$PL_2(S(x,i),x'(j)) = \max_{(y,y') \in Y(x,x'(j)), y \in S(x,i)} \sum_{x'' \in W((x,x'(j)),(y,y'))} C(S(x,i),x'').$$

That is, the second component plausibility factor is defined as the maximum plausibility factor of alignments that suggest the sense. The plausibility factor of each alignment is defined as the sum of correlations between the sense and the translingually alignable accompanying words.

[Formula II]

$$PL_2(S(x,i),x'(j)) = \max_{(y,y') \in Y(x,x'(j)), y \in S(x,i)} MI(y,y') \cdot \sum_{x'' \in W((x,x'(j)),(y,y'))} C(S(x,i),x'').$$

In this alternative, the plausibility factor of each alignment is defined as the product of the mutual information of the second-language word association and the sum of

correlations between the sense and the translingually alignable accompanying words.

The above definition of the correlations between senses and clues is recursive, so we calculate them iteratively with the following initial values:

$$C_0(S(x, i), x'(j)) = MI(x, x'(j)).$$

That is, the mutual information value of the target word and a clue is used as the initial values for the correlations between all the senses and the clue. The number of iterations needed was determined experimentally together with the value of parameter α used to adjust the relative weights of the component plausibility factors.

4.3.6 Example of convergence of sense-vs.-clue correlations

The sense-vs.-clue correlation values converge within several iterations, as shown in Fig. 4.5. The examples shown are the results of using Formula II with $\alpha = 5$ in the experiment described in Section 4.5. The curves show the change in correlation values between two senses of "tank," {tank, タンク<*TANKU*>, 水槽<*SUISOU*>, 槽<*SOU*>} and {tank, 戦車<*SENSHA*>}, and four clues, "troop," "ozone," "Poland," and "safety."

The curves for "troop" show a typical pattern—while the correlation with the relevant sense keeps the initial value, those with irrelevant senses decrease as the iterations proceed.

The curves for "ozone" show another pattern—the correlation value(s) with irrelevant sense(s) begin to decrease at the second cycle. This pattern is specific to the case in which a first-language word association is not aligned with a second-language one. In this case, the divergence in correlation values is caused by the difference in correlation values between the senses and the accompanying words. Therefore, it occurs one cycle behind the divergence in correlation values between the senses and the accompanying words. The English word association (tank, ozone) was not aligned with any Japanese word association, so the correlations between the senses and "ozone" were calculated based only on the set of accompanying words regardless of translingual alignability, {air, area, car, control, deep, defense, emission, fuel, gas, gasoline, pump, road, study, upper, vapor}. The majority of these accompanying words had larger correlations with {tank, タンク<*TANKU*>, 水槽<*SUISOU*>, 槽<*SOU*>} than with {tank, 戦車<*SENSHA*>}. Consequently, "ozone" had a larger correlation with {tank, タンク<*TANKU*>, 水槽<*SUISOU*>, 槽<*SOU*>} than with {tank, 戦車<*SENSHA*>}.

The curves for "Poland" demonstrate that correlations between senses and clues can be calculated correctly even when a first-language word association is aligned only with incorrect second-language word association(s). The English word association (tank, Poland) was aligned with the Japanese word association (水槽<*SUISOU*>, 波<*NAMI*>),

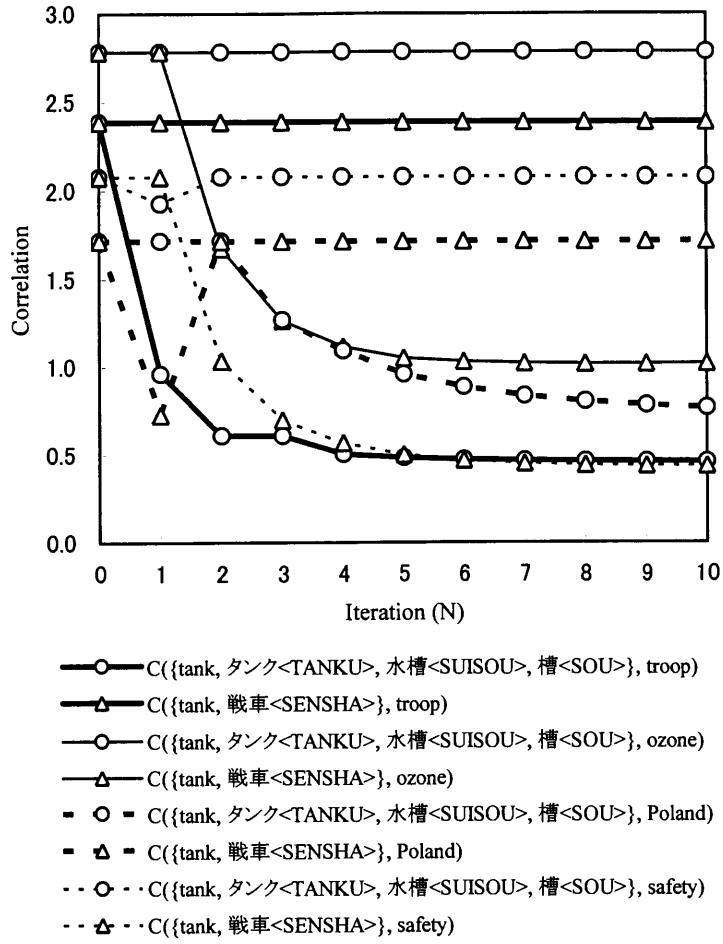


Figure 4.5 Convergence of sense-vs.-clue correlations

and the alignment was characterized by a set of one translingually alignable accompanying word, {government}. This set was far smaller than the set of accompanying words regardless of translingual alignability, {Army, Belarus, GM, German, NATO, Polish, Russia, Russian, World War, car, economy, government, parliament, treaty}, the majority of which had larger correlations with {tank, 戦車 <*SENSHA*>} than with {tank, タンク<*TANKU*>, 水槽<*SUISOU*>, 槽<*SOU*>}. Consequently, despite the alignment of (tank, Poland) with (水槽<*SUISOU*>, 波<*NAMI*>), "Poland" had a larger correlation with {tank, 戦車<*SENSHA*>} than with {tank, タンク<*TANKU*>, 水槽 <*SUISOU*>, 槽<*SOU*>}.

The curves for "safety" demonstrate that a larger set of accompanying words regardless of translingual alignability compensates for the deficiency of small sets of translingually alignable accompanying words. The English word association (tank,

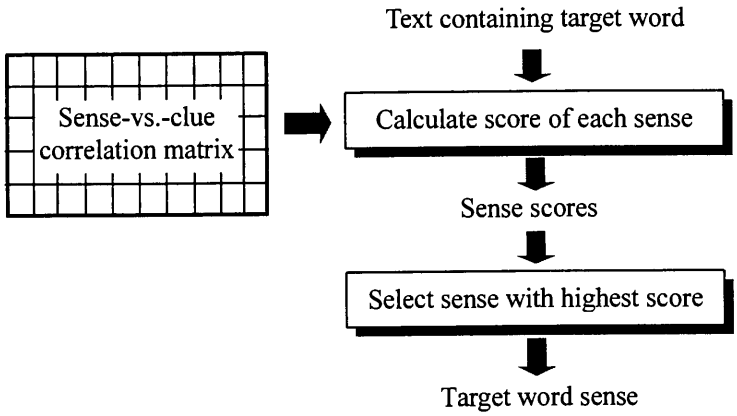


Figure 4.6 Flow of word sense disambiguation using sense-vs.-clue correlation matrix

safety) was aligned with two Japanese word associations (タンク<*TANKU*>, 安全<*ANZEN*>) and (戦車<*TANKU*>, 安全<*ANZEN*>). Both alignments were characterized by small sets of translingually alignable accompanying words, {car, design, fuel, truck, vehicle} and {car, government, study, vehicle}. Therefore, the correlation values were unstable in the first cycle. However, in the second cycle, they began to converge due to a relatively large set of accompanying words regardless of translingual alignability, {FDA, Ford, GM, Pena, Secretary, Transportation, air, car, design, fire, fuel, fuel tank, government, jet, owner, pickup, recall, study, truck, vehicle}.

4.4 Word sense disambiguation using sense-vs.-clue correlation matrix

The sense-vs.-clue correlation calculation method is difficult to evaluate directly through its output correlation matrices. We therefore evaluated the method through a WSD experiment using sense-vs.-clue correlation matrices. WSD is one of the most important applications of word sense association networks. We describe below our method for WSD using sense-vs.-clue correlation matrices.

As illustrated in Fig. 4.6, for each instance of a target word, the sense scores are calculated based on the clues in the context, and then the sense that maximizes the score is selected. When two or more senses maximize the score, neither is selected. We have the following alternative formulae for defining the score. In them, $Score(S(x_0, i))$ denotes the score of the i-th sense, $S(x_0, i)$, of target word x_0, and x_p denotes the word whose position relative to x_0 is p. The value of p is negative for words preceding x_0, and it is positive for words following x_0. The context examined is a window consisting of

(γ+1) words centered on x_0. We experimentally identified the best formula together with the optimum value of γ.

$$[\text{Formula A}]\ \text{Score}(S(x_0,i)) = \sum_{1\le|p|\le\frac{\gamma}{2}} \frac{1}{\sqrt{|p|}} \cdot C(S(x_0,i),x_p)$$

$$[\text{Formula B}]\ \text{Score}(S(x_0,i)) = \sum_{1\le|p|\le\frac{\gamma}{2}} \frac{1}{\sqrt[4]{|p|}} \cdot C(S(x_0,i),x_p)$$

$$[\text{Formula C}]\ \text{Score}(S(x_0,i)) = \sum_{1\le|p|\le\frac{\gamma}{2}} C(S(x_0,i),x_p)$$

$$[\text{Formula D}]\ \text{Score}(S(x_0,i)) = \max_{1\le|p|\le\frac{\gamma}{2}} C(S(x_0,i),x_p)$$

Formulae A and B are weighted sums of the correlations between the sense and clues in the context. Both formulae weight clues to reflect their distances from the target word; Formula A reduces the weight more quickly with the distance than Formula B. Formula C is the sum of correlations between the sense and clues in the context. Formula D is the maximum of the correlations between the sense and clues in the context.

Two examples of determining the sense of target word "tank" are shown in Fig. 4.7. The scores of the senses were calculated according to Formula C, with γ set to 50. The senses were determined correctly in both examples.

4.5 Experimental evaluation

4.5.1 Method and materials

We evaluated our method experimentally using corpora of English and Japanese newspaper articles. The first language was English, and the second was Japanese. A Wall Street Journal corpus (July 1994 to December 1995; 189 MB) and a Nihon Keizai Shimbun corpus (December 1993 to November 1994; 275 MB) were used as the comparable corpus for training. To extract as many word associations as possible from the corpus of each language, we set the thresholds very low:

- threshold for occurrence frequencies of words: 10
- threshold for mutual information: 0.0.

These settings were common to both corpora.

A bilingual dictionary was prepared by collecting pairs of nouns that are translations of one another from the EDR English-to-Japanese and Japanese-to-English dictionaries (EDR 1990b). The resulting dictionary, a collection of 633,000 pairs of 269,000 English

...searched a Bowmar Instrument Corp. unit yesterday as part of a criminal investigation of allegedly substandard testing involving electronic components for the Army's M1A2 Abrams *tank* and Patriot missile systems, according to people familiar with the case. Agents from the Defense Criminal Investigative Service, an arm of the Pentagon's inspector general,....

Score(*S1*={tank, タンク<*TANKU*>, 水槽<*SUISOU*>, 槽<*SOU*>})
= *C*(*S1*, Army)+*C*(*S1*, missile)+*C*(*S1*, Defense)
= 0.551+0.061+0.008 = 0.620

Score(*S2*={tank, 戦車<*SENSHA*>})
= *C*(*S2*, Army)+*C*(*S2*, missile)+*C*(*S2*, Defense)
= 1.823+2.278+1.089 = 5.190

⇒ Sense: {tank, 戦車<*SENSHA*>}

(a) Example 1 (Wall Street Journal 1996/4/26)

...rising price of gas by raising their prices accordingly. Alamo, for example, is charging rising local prices to customers who want to prepay for a *tank* of gas. It also hiked the penalty for returning your car with an empty tank to $3.25 a gallon, from $2.99 a gallon. "It's not....

Score(*S1*={tank, タンク<*TANKU*>, 水槽<*SUISOU*>, 槽<*SOU*>})
= *C*(*S1*, gas)+*C*(*S1*, gas)+*C*(*S1*, car)+*C*(*S1*, gallon)+*C*(*S1*, gallon)
= 1.077+1.077+0.822+2.006+2.006 = 6.988

Score(*S2*={tank, 戦車<*SENSHA*>})
= *C*(*S2*, gas)+*C*(*S2*, gas)+*C*(*S2*, car)+*C*(*S2*, gallon)+*C*(*S2*, gallon)
= 0.021+0.021+0.054+0.526+0.526 = 1.148

⇒ Sense: {tank, タンク<*TANKU*>, 水槽<*SUISOU*>, 槽<*SOU*>}

(b) Example 2 (Wall Street Journal 1996/5/10)

[Note] The target word is italicized, and clues are underlined in the text.

Figure 4.7 Examples of sense determination

nouns and 276,000 Japanese nouns, includes many unusual pairs of translation equivalents. While its size eases the problem of failure in alignment, the huge number of possible pairs of translation equivalents worsens the problem of ambiguity in alignment. Although our method addresses both problems, we prefer easing the problem of failure in alignment.

We selected 60 English polysemous nouns as the test target words. Words appearing

in newspapers and with different senses were chosen. The frequencies of the test words in the training corpus ranged from 39,140 ("share," the third noun in descending order of frequency) to 106 ("appreciation," the 2,914th noun), and the median was 410. The senses of each test word were defined manually. The senses were rather coarse-grained; i.e., they basically corresponded to groups of translation equivalents within the entries of everyday English-Japanese dictionaries. The number of senses per test word ranged from 2 to 8, and the average was 3.4. The average number of clues per test word, i.e., word associations in which each test word was involved, was 175. For each test word, a sense-vs.-clue correlation matrix was calculated using Method I and Method II, which define the second component plausibility factor using Formula I and Formula II, respectively (See Subsection 4.3.5).

For evaluation, 100 passages per test word were selected randomly from a Wall Street Journal corpus (January to December 1996) with a publishing period different from that of the training corpus. The occurrence frequencies of three test words in the test corpus were less than 100, i.e., 71, 82, and 82, so the total number of passages was 5,935. The sense of the test word in each passage was determined by using Formulae A, B, C, and D (See Section 4.4), and the results were compared with human-judged senses.

We used applicability and precision to evaluate the WSD performance (Dagan and Itai 1994). Applicability (A_{WSD}) is the proportion of instances of the test word(s) that the method could disambiguate. Precision (P_{WSD}) is the proportion of disambiguated instances of the test word(s) that the method disambiguated correctly. Although applicability and precision are easy to understand, they need be interpreted pair-wise.

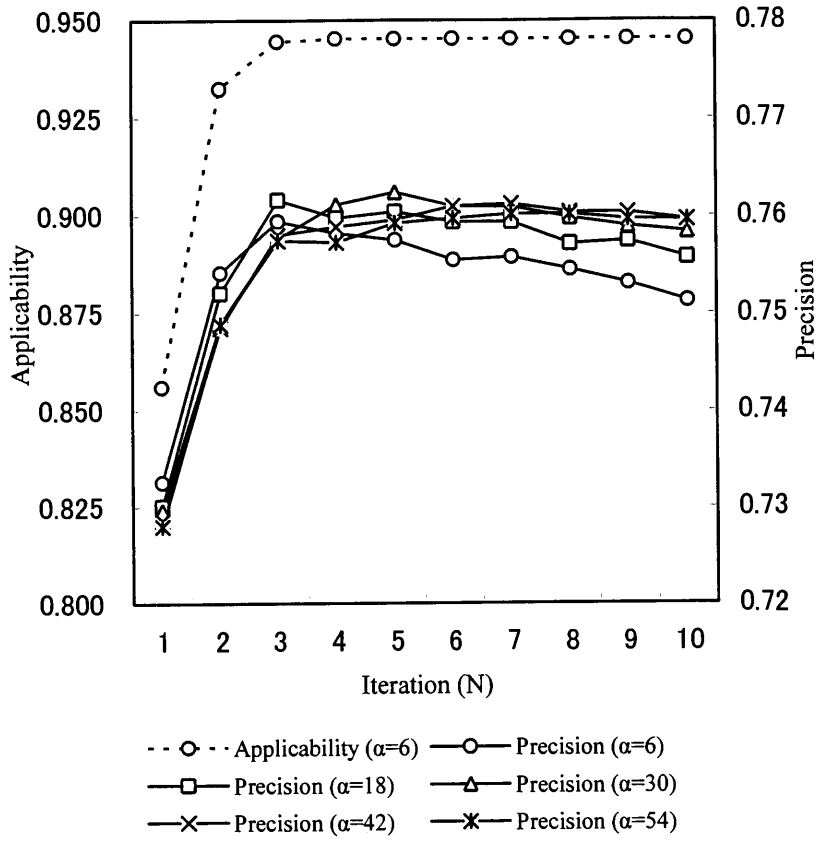
We also calculated the F-measure, since single measurement is preferable for determining the optimum values of the parameters. The F-measure is defined as the harmonic means of recall (R_{WSD}) and precision (van Rijsbergen 1979):

$$F_{WSD} = \frac{2 \cdot R_{WSD} \cdot P_{WSD}}{R_{WSD} + P_{WSD}}.$$

Recall is the proportion of instances of test word(s) that the method disambiguated correctly. Substituting $R_{WSD} = A_{WSD} \cdot P_{WSD}$ for R_{WSD}, we get

$$F_{WSD} = \frac{2 \cdot A_{WSD} \cdot P_{WSD}}{1 + A_{WSD}}.$$

In our experimental evaluation, we calculated applicability, precision, and the F-measure for each test word and calculated the average for the 60 test words. Unless otherwise stated, the average values are used in the following.



[Note] Formula A ($\gamma = 100$) was used for WSD.

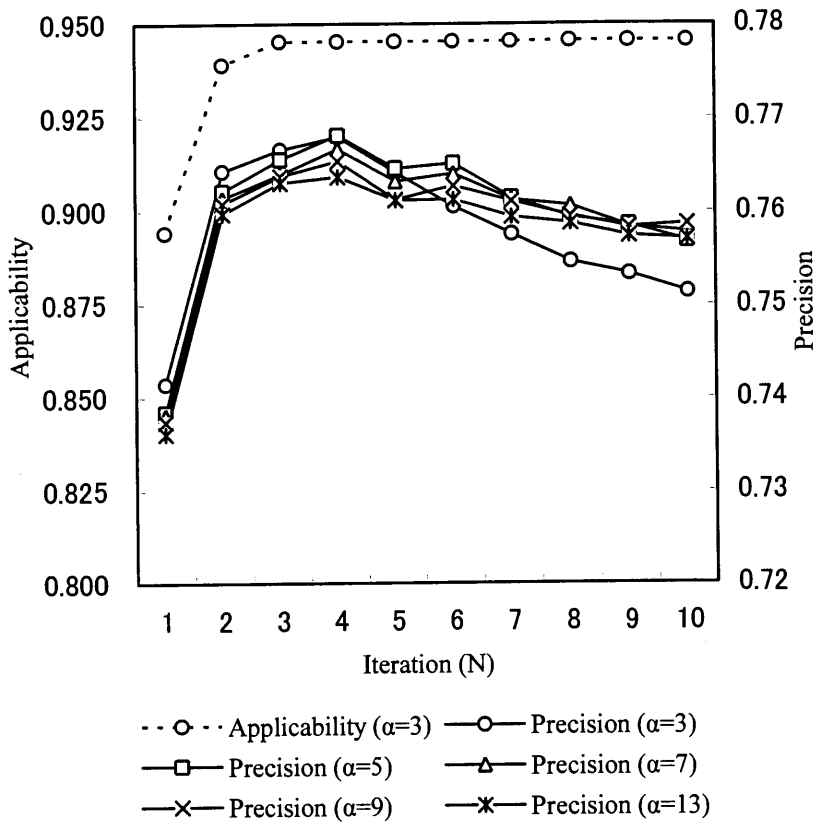
Figure 4.8 WSD performance using sense-vs.-clue correlation matrices produced by Method I

4.5.2 Comparison of formulae for defining second component plausibility factor

Sense-vs.-clue correlation matrices were calculated using Methods I and II with various parameter settings. Then, the applicability and precision of WSD using the matrices were calculated. Figures 4.8 and 4.9 show WSD performance when we used Formula A with $\gamma = 100$. The curves show how the applicability and precision varied with the number of iterations with parameter α fixed.

Applicability was not sensitive to α, so only the results for $\alpha = 6$ (Method I, Fig. 4.8) and $\alpha = 3$ (Method II, Fig. 4.9) are shown. In both methods, applicability increased with N and saturated at $N \geq 3$.

Precision was sensitive to α, and the results for $\alpha = 6, 18, 30, 42, 54$ (Fig. 4.8) and $\alpha = 3, 5, 7, 9, 13$ (Fig. 4.9) are shown. Precision first increased with N, and then decreased. It decreased with large N probably because clues whose correlations with senses



[Note] Formula A (γ = 100) was used for WSD.

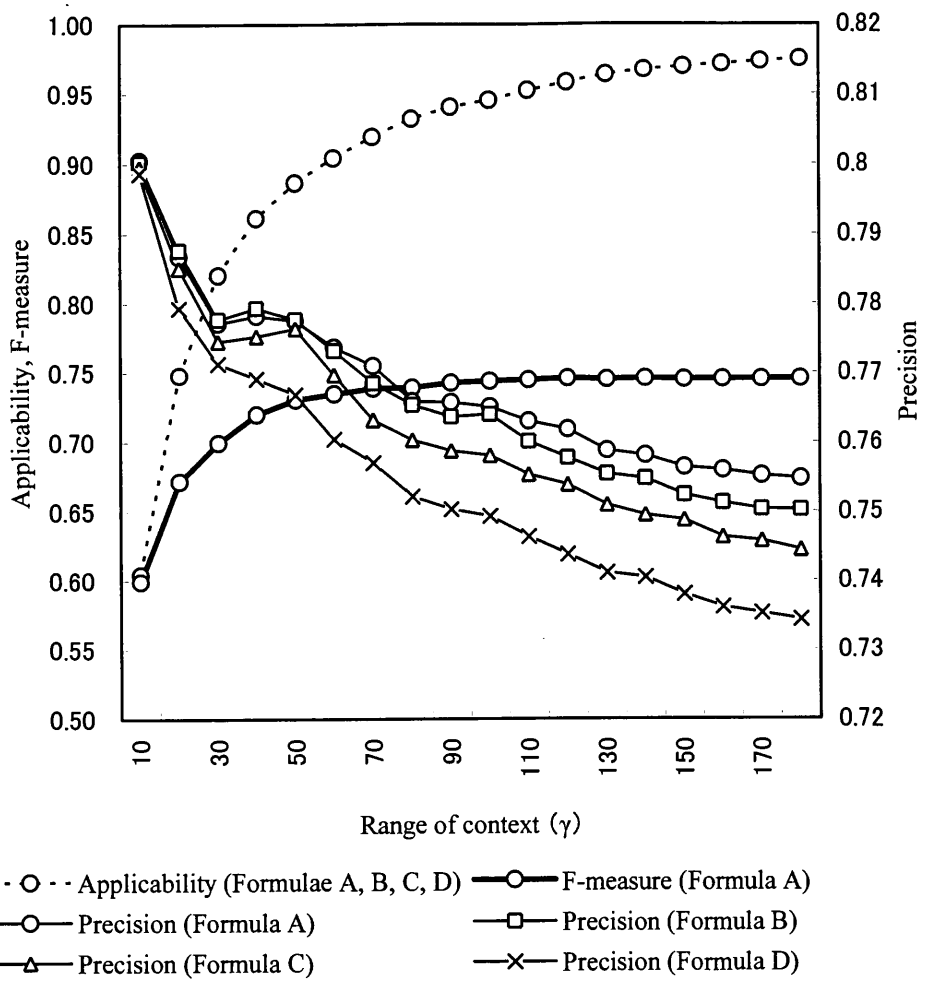
Figure 4.9 WSD performance using sense-vs.-clue correlation matrices produced by Method II

converge slowly are usually less reliable than those whose correlations converge quickly. The peak value of precision and the range of N in which the peak value was maintained depended on α. When α was too small, the precision began to decrease immediately after reaching a peak. When α was too large, the peak was not very high, although it was maintained over a relatively wide range of N.

The optimum ranges of the parameter values proved to be rather wide:

- Method I: $24 \leqq \alpha \leqq 36$, $4 \leqq N \leqq 8$
- Method II: $5 \leqq \alpha \leqq 7$, $3 \leqq N \leqq 6$

Method II was more sensitive to N than Method I; both applicability and precision increased quickly, but precision decreased quickly after reaching its peak. When the parameters were set appropriately, Method II worked slightly better than Method I. That is, the peak precision with Method II was about 0.5% higher than with Method I, and the applicability was almost the same. However, the difference is insignificant.



[Note] Method II ($\alpha = 5$, $N = 6$) was used to produce the sense-vs.-clue correlation matrix.

Figure 4.10 WSD performance using different formulae for defining the score

4.5.3 Comparison of formulae for defining sense score

Formulae A, B, C, and D were compared experimentally using the same sense-vs.-clue correlation matrix. Figure 4.10 shows the WSD performance for Method II with $\alpha = 5$ and $N = 6$. The curves show how the applicability, precision, and F-measure varied with the context range, γ. The applicability curve is common to all four formulae. The F-measure curve is shown only for Formula A.

The results show that weighting the clues to reflect their distances from the target word is effective and that the weight should be reduced relatively quickly with the distance. They also show that a relatively wide range of context should be used.

Moreover, there is a trade-off between applicability and precision: applicability increased with γ, while precision decreased. Consequently, the *F*-measure was at maximum when γ was between 100 and 180.

In sum, experiments showed that Method II combined with Formula A is the best combination. The optimum parameter ranges were $5 \leqq \alpha \leqq 7$, $3 \leqq N \leqq 6$, and $100 \leqq \gamma \leqq 180$. With $\alpha = 5$, $N = 6$, and $\gamma = 120$, Method II combined with Formula A achieved applicability of 95.8%, precision of 76.2%, and an *F*-measure of 74.6%.

4.5.4 Detailed evaluation of word sense disambiguation results

The performance of our method for 6 of the 60 test words, i.e., "measure," "promotion," "race," "tank," "title," and "trial," is summarized in Table 4.2. The incidence matrices show the results of experiments using Method II with $\alpha = 5$ and $N = 6$ and Formula A with $\gamma = 120$. The rows and columns of the matrices represent human-judged correct senses and senses determined by our method, respectively, and each cell shows the number of test passages or instances of the test word.

Applicability and precision, especially precision, varied by test word. Our method performed fairly well for the frequent senses, but not so well for the infrequent ones. It identified topic-specific senses well, but not generic senses. The poor performance for "measure" (Table 4.2(a)) is explained as follows. The second sense of "measure," {measure, 対策<*TAISAKU*>, 手段<*SHUDAN*>, 処置<*SHOCHI*>}, is a very generic sense. Therefore, effective clues identifying that sense could not be acquired.

The poor performance for "race" (Table 4.2(c)) indicates the limitation of the "one sense per collocation" hypothesis. The first sense of "race," {race, レース<*REESU*>, 競争<*KYOUSOU*>, 競走<*KYOUSOU*>, 争い<*ARASOI*>, 戦<*SEN*>}, is correlated with a topic, "race for the presidency," and the second sense of "race," {race, 人種<*ZINSHU*>, 民族<*MINZOKU*>, 種属<*SHUZOKU*>}, is correlated with another topic, "racial discrimination," and both topics are related to a broader topic, "politics." Therefore, many clues were shared by these two senses.

Although much work has been done on word sense disambiguation, researchers, including us, have evaluated their methods by using their own sense inventory, training corpus, and test corpus. This makes it difficult to compare one method with another method directly. Therefore, we compared our method with a method that always selects the most frequent sense, independently of context. This most-frequent sense selection method is suitable as a baseline because its performance reflects the ratios of the senses of the target word, i.e., the difficulty of the task (Gale, et al. 1992a; Dagan and Itai 1994).

Table 4.2 Results of word sense disambiguation for six test words

(a) "measure" (Applicability=99%; Precision=47.5%)

Correct sense \ Results	S1	S2	S3	?	Total
S1={measure, 寸法, 大きさ, 量, 程度, 尺度, 指数, 基準, 測定, 計測}	22	0	14	1	37
S2={measure, 対策, 手段, 処置}	6	0	32	0	38
S3={measure, 法案, 議案, 法令}	0	0	25	0	25
Total	28	0	71	1	100

S1: a system or instrument for calculating amount, size, weight, etc.
S2: an action taken to gain a certain end
S3: a law suggested in Parliament

(b) "promotion" (Applicability=99%; Precision=85.9%)

Correct sense \ Results	S1	S2	S3	?	Total
S1={promotion, 宣伝, 売り込み, 販売促進, プロモーション}	72	1	0	0	73
S2={promotion, 昇格, 昇進, 昇任, 就任, 登用, 進級}	10	13	0	1	24
S3={promotion, 奨励, 振興, 促進, 増進, 助長}	2	1	0	0	3
Total	84	15	0	1	100

S1: an activity intended to help sell a product
S2: advancement in rank or position
S3: action to help something develop or succeed

(c) "race" (Applicability=90%; Precision=52.2%)

Correct sense \ Results	S1	S2	S3	?	Total
S1={race, レース, 競争, 競走, 争い, 戦}	29	40	0	7	76
S2={race, 人種, 民族, 種属}	3	18	0	3	24
S3={race, 水路, 用水}	0	0	0	0	0
Total	32	58	0	10	100

S1: any competition, or a contest of speed
S2: one of the groups that people can be divided into according to physical features, history, language, etc.
S3: a channel for a current of water

(d) "tank" (Applicability=99%; Precision=86.9%)

Correct sense \ Results	S1	S2	?	Total
S1={tank, タンク, 水槽, 槽}	60	4	0	64
S2={tank, 戦車}	9	26	1	36
Total	69	30	1	100

S1: a large container for storing liquid or gas
S2: an enclosed heavily armed, armored vehicle

(e) "title" (Applicability=98%; Precision=86.7%)

Correct sense \ Results	S1	S2	S3	S4	?	Total
S1={title, 肩書き, 称号, 敬称}	45	1	0	0	0	46
S2={title, 題名, 題目, 表題, 書名}	2	34	0	0	2	38
S3={title, 権利, 資格, 所有権}	3	0	0	1	0	4
S4={title, 選手権}	5	1	0	6	0	12
Total	55	36	0	7	2	100

S1: a word or name given to a person to be used before his/her name as a sign of rank, profession, etc.
S2: a name given to a book, play, etc.
S3: the legal right to own something
S4: the position of being the winner of a sports competition

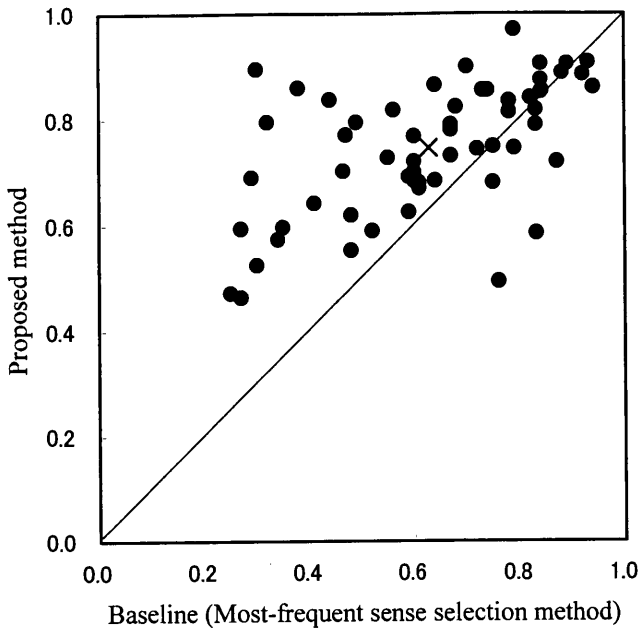
(f) "trial" (Applicability=98%; Precision=90.8%)

Correct sense \ Results	S1	S2	S3	S4	S5	?	Total
S1={trial, 裁判, 公判, 審理}	66	2	0	0	0	2	70
S2={trial, 試し, 試み, 試験, 実験, 試用}	6	23	0	0	0	0	29
S3={trial, 予選}	1	0	0	0	0	0	1
S4={trial, 厄介, 困り者}	0	0	0	0	0	0	0
S5={trial, 苦難, 試練, 辛苦}	0	0	0	0	0	0	0
Total	73	25	0	0	0	2	100

S1: a legal process in which a court examines a case
S2: a process of testing to determine quality, value, usefulness, etc.
S3: a sports competition that tests a player's ability
S4: annoying thing or person
S5: difficulties and troubles

[Note 1] Method II (α=5, N=6) was used to produce the sense-vs.-clue correlation matrices, and Formula A (γ=120) was used for WSD.

[Note 2] The columns labeled by the question mark represent "inapplicable" cases.



[Note] ● test words, × average

Figure 4.11 Distribution of *F*-measures: proposed method vs. baseline method

To implement the most-frequent sense selection method, we had to determine which sense of each target word occurs most frequently. This is not trivial because the training corpus is not sense-tagged. Therefore, we implemented a variation of our method to determine the most frequent senses. It calculates the sum of the correlations for each row of a sense-vs.-clue correlation matrix and takes the sense corresponding to the row with the maximum value as the most frequent one. This is a simple but effective method for determining the most frequent sense because the number of clues identifying the sense is strongly correlated with the frequency of the sense.

The most-frequent sense selection method was tested using the sense-vs.-clue correlation matrices obtained by Method II with $\alpha = 5$ and $N = 6$. Its precision was 62.8%, averaged over the 60 test target words.[7] Note that the applicability of the most-frequent sense selection method is always 100%, so the *F*-measure is the same as the precision. The *F*-measures our method and the most-frequent sense selection method achieved for the 60 test target words are compared in Fig. 4.11. More target words were

[7] The most-frequent sense selection method does not always select the truly most-frequent senses. Although the most-frequent senses in the training corpus do not always coincide with those in the test corpus, selecting senses that are most frequent in the test corpus resulted in precision of 66.8%. This is considered to be the upper limit of precision of the most-frequent sense selection method.

plotted to the upper left of the diagonal, indicating that our method outperformed the baseline one.

Finally, we mention the combination of the most-frequent sense selection method with our method. Most applications of WSD require 100% applicability. A simple method to meet this requirement is to switch to the most-frequent sense selection method when our method cannot determine the sense. The precision of this combined method was experimentally shown to be about 75%.

4.5.5 Comparison with alternative methods

Two alternative sense-vs.-clue correlation calculation methods were compared with our method to evaluate the effectiveness of the set of accompanying words regardless of translingual alignability (Subsection 4.2.3) and the set of translingually alignable accompanying words (Subsection 4.2.2). The performance of WSD using sense-vs.-clue correlation matrices produced by Method II and these alternative methods is shown in Fig. 4.12.

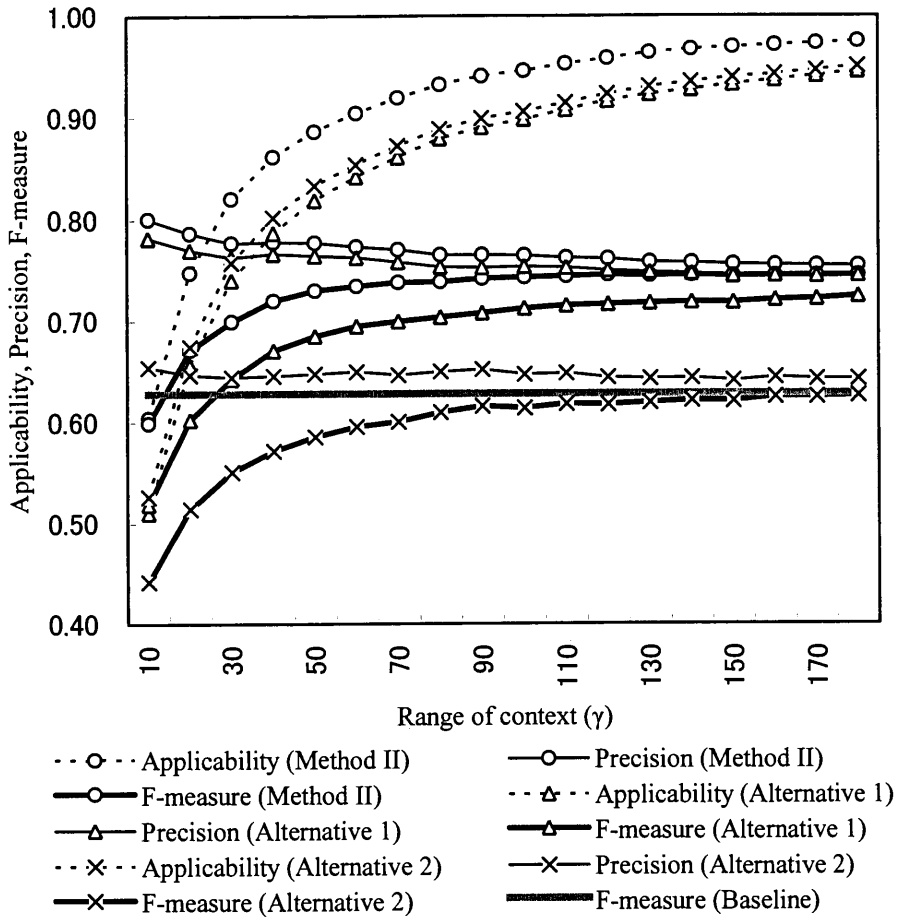
Alternative 1 is a variant of our method that uses only the set of translingually alignable accompanying words. That is, the formula for defining the plausibility factor in Subsection 4.3.5 is replaced with

$$PL(S(x,i),x'(j)) = PL_2(S(x,i),x'(j)).$$

Formula II is used for the second component plausibility factor, PL_2.

Figure 4.12 shows that Method II achieved much higher applicability and a little higher precision than Alternative 1. The reason for the improved applicability is that failure in word association alignment is effectively remedied by the iterative calculation process using the sets of accompanying words regardless of translingual alignability. Inspecting the data for the 60 test target words revealed that the rate of success in aligning word associations was only 42.9%. Precision was also improved by using the sets of accompanying words regardless of translingual alignability. This means that the usefulness of clues does not depend on whether they can be aligned with counterparts in another language. Thus, iterative calculation using sets of accompanying words regardless of translingual alignability is essential in our method.

Alternative 2 uses neither the set of translingually alignable accompanying words nor the set of accompanying words regardless of translingual alignability. It is based on the assumption that alignments with strong word associations are preferable to those with weak word associations. The correlation between the i-th sense $S(x, i)$ of the target word x and the j-th clue $x'(j)$ is defined as



[Note] Parameters were set as α=5 and N=6 for Method II, and N=6 for alternative 1. Formula A was used for WSD.

Figure 4.12 WSD performance using sense-vs.-clue correlation matrices produced by Method II and alternatives 1 and 2

$$C(S(x,i),x'(j)) = \max_{(y,y')\in Y(x,x'(j)),y\in S(x,i)} MI(y,y'),$$

where $Y(x, x'(j))$ is the set of possible counterparts of first-language word association (x, $x'(j)$), and $MI(y, y')$ is the mutual information of second-language word association (y, y').

Figure 4.12 shows that Alternative 1 achieved much higher precision and a little lower applicability than Alternative 2. The difference in precision reflects the reliability of the underling assumptions: the assumption that word association alignments accompanied by many translingually alignable words are preferable is reliable, but the assumption that alignments with strong word associations are preferable is not reliable.

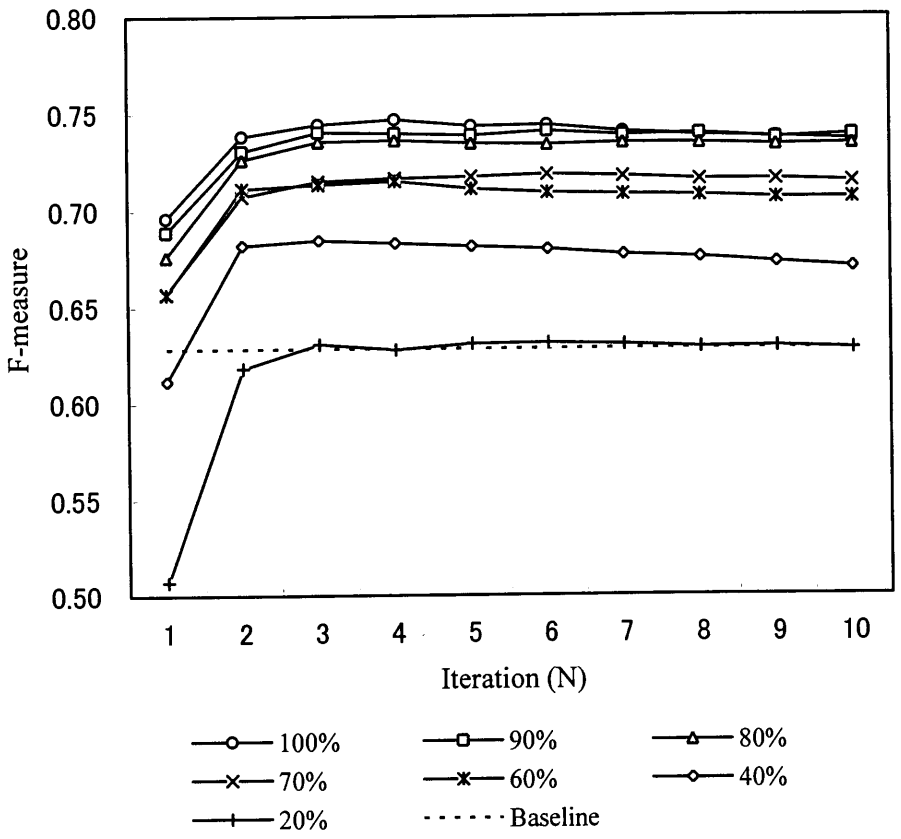
The reason Alternative 1 has a slightly lower applicability is that it neglects word associations without accompanying words.

4.5.6 Sensitivity to bilingual dictionary and corpus

Obviously, the performance of our method depends on the bilingual dictionary used to align word associations as well as to align accompanying words. The sensitivity of the performance to the bilingual dictionary was evaluated as follows. First, a series of reduced bilingual dictionaries were produced by deleting randomly selected pairs of translation equivalents from the bilingual dictionary described in Subsection 4.5.1. Then, sense-vs.-clue correlation matrices were calculated using these reduced dictionaries. Finally, WSD experiments were done using the resulting sense-vs.-clue correlation matrices.

Figure 4.13 shows the results of the experiments using the dictionaries reduced to 90%, 80%, 70%, 60%, 40%, and 20%. The curves show how the *F*-measure varied with the number of iterations in the calculation of the sense-vs.-clue correlation matrix. We see that the *F*-measure did not greatly decrease until the dictionary was reduced to 80%, and it remained higher than that of the baseline (62.8%) even when the dictionary was reduced to 40%. Thus, our method is workable with incomplete bilingual dictionaries owing to the iterative algorithm. It should be noted that the reduced bilingual dictionaries, which were produced by deleting a certain percentage of entries whether they were frequent ones or not, were much more incomplete than ordinary bilingual dictionaries of the same size.

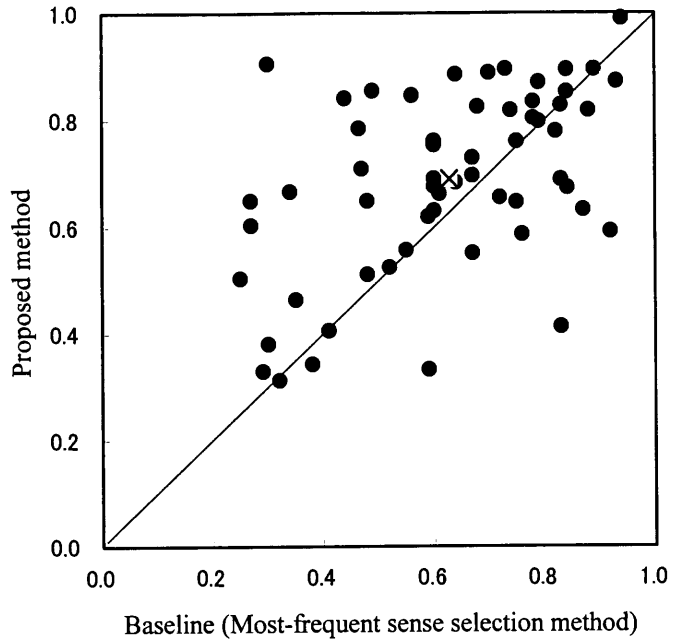
The performance of our method also varied with the comparability of the training corpus. The combination of Wall Street Journal (WSJ) and Nihon Keizai Shimbun (Nikkei) corpora is a difficult one. Both focus on economic and political news. However, their contents are quite different, since the majority of articles are related to domestic events in their countries, the U.S. and Japan. Given the disparity between the WSJ and Nikkei corpora, the fairly good results described above are rather surprising. More tightly comparable corpora, e.g., a combination of The Daily Yomiuri and The Yomiuri Shimbun, which are published by the same company, would result in better performance.



[Note] Method II ($\alpha = 5$) was used to produce sense-vs.-clue correlation matrices, and Formula A ($\gamma = 100$) was used for WSD.

Figure 4.13 WSD performance using sense-vs.-clue correlation matrices produced using reduced dictionaries

More importantly, the bound on comparability of corpora the present method accepts should be clarified. We did an additional experiment using the WSJ corpus and a Mainichi Shimbun corpus. The Mainichi Shimbun is a general paper and less similar to the WSJ than the Nikkei; it covers a wide range of subjects including politics, economy, sports, culture, and local news. The Mainichi corpus consisted of articles from January to December 1994, and the total size was 142 MB. The experiment was done by using the same method, parameter values, test target words, and test passages as that with the WSJ and Nikkei described in Subsection 4.5.4. The distribution of *F*-measures for the 60 target words is shown in Fig. 4.14. The average *F*-measure was 69.0%. The performance for the WSJ and Mainichi was much poorer compared to that for the WSJ and Nikkei (Fig. 4.11). The combination of the Wall Street Journal and Mainichi



[Note] ● test words, × average

Figure 4.14 Results of WSD using Wall Street Journal and Mainichi Shimbun corpora

Shimbun thus proved to be beyond the bound on acceptable comparability.

4.6 Discussion

4.6.1 Strong and weak points of proposed method

Our method differs from conventional WSD methods in an important way. The conventional methods deal with a large number of tokens (instances) of the target word along with their contexts individually (Black 1988; Zernik 1991; Hearst 1991; Yarowsky 1992; Luk 1995; Yarowsky 1995; Karov and Edelman 1998; Schuetze 1998). In contrast, our method does not deal with individual tokens of the target word in the training corpus. It learns from a collection of word associations (to be unambiguous, associations between types) into which the contexts of all tokens of the target word are condensed. The reason for doing so is that comparable corpora do not contain correspondence between tokens, unlike parallel corpora. In addition, our method uses not only associations in which the target word is present but also associations between words associated with the target word. These word associations are extracted from the whole corpus; it is not limited to the contexts of the target word.

These characteristics strengthen our method. First, it is effective even for target words with a relatively small number of examples. That is, clues can be acquired reliably together with associated clues, even if they do not co-occur with the target word very frequently. In the experiment described in Section 4.5, the number of training examples per target word was small compared to those in most of the other work. The median frequency of the 60 target words in the training corpus was 410; i.e., 30 target words had 410 or less examples. Second, the method is not computationally hard, although it calculates the sense-vs.-clue correlations iteratively. It manipulates a relatively small matrix, i.e., several senses by a few hundred clues, for each target word. Moreover, the correlations converge rapidly because the clues are densely associated with each other. It should be added that the large amount of computation to extract word associations from the training corpus is shared by all target words.

The above characteristics also lead to a weakness in our method—an error occurs when the "one sense per collocation" hypothesis does not hold. For example, "Republican" is selected as a clue for the target word "race," as they are strongly associated with each other. As a result of the iterative calculation, "Republican" has high correlation with one and only one sense of "race." This is inappropriate given that "Republican" actually occurs in the context of racial discrimination as well as in the context of a presidential race. This problem sometimes is serious because the error can propagate through the iterative calculation process.

The sense-vs.-clue correlation calculation method, which is characterized by its iterative algorithm, looks like the expectation maximization (EM) algorithm (Dempster, et al. 1977). However, it is not the EM algorithm. Note that we cannot observe data, e.g., pairs of corresponding contexts of two languages, to estimate the parameters of a probabilistic model. All we can use are collections of word associations of the two languages. The method is based on the heuristics that the correlation between a sense and a clue depends on the plausibility of word association alignments suggesting the sense and the clue, which depends on the correlations between the sense and accompanying clues. This circularity results in the iterative algorithm. Although the algorithm does not have a solid mathematical foundation, as does the EM algorithm, the experiment demonstrated that the correlations converge rapidly.

4.6.2 Limitations and directions for extension

Although it has produced promising results, the developed method has a few problems. These limitations, along with future extensions, are discussed below.

(1) Multilingual distinction of senses

The developed method is based on the premise that the senses of a polysemous word in a language are lexicalized differently in another language. However, this is not always true; that is, the ambiguity of a word may be preserved by its translation equivalents. As described in Subsection 4.3.2, it is better to use translation equivalents that do not preserve the ambiguity. However, doing so is useless unless they are frequently used translation equivalents. A promising approach to solving this problem is to use two or more second languages (Resnik and Yarowsky 2000).

(2) Use of syntactic relations

The developed method extracts word associations based on co-occurrence in a medium-sized window. In other words, it uses topically related words as clues for disambiguation. However, it is commonly accepted that different types of clues are required for disambiguation, or different types of clues are appropriate for different kinds of words (Ide and Veronis 1998). Particularly, syntactically related words are more useful for some kinds of polysemous words.

It is an important and interesting research issue to extend our method so that it can acquire clues based on syntactic co-occurrence. The framework of the method is compatible with syntactic co-occurrence. A parser for the first language is indispensable, while a parser for the second language is dispensable. For the second language, we can use co-occurrence in a small window instead of syntactic co-occurrence.

We expect that the performance of WSD will be improved by using syntactically related clues. For example, the sense of "measure" is judged to be {measure, 対策 *<TAISAKU>*, 手段*<SHUDAN>*, 処置*<SHOCHI>*} ("an action taken to gain a certain end") when it is the object of "take," while the sense of "measure" is judged to be {measure, 法案*<HOUAN>*, 議案*<GIAN>*, 法令*<HOUREI>*} ("a law suggested in Parliament") when it is the object of "pass," "vote," or "approve." Likewise, the sense of "race" is judged to be {race, レース*<REESU>*, 競争*<KYOUSOU>*, 競走*<KYOUSOU>*, 争い*<ARASOI>*, 戦 *<SEN>*} ("any competition, or a contest of speed"), when it is the object of "win" or "lose," or it is modified by "presidential," "congressional," "tight," or "tough."

Finally, we mention sense disambiguation of polysemous verbs. Usually, the object and subject of a verb as well as nouns in prepositional phrases attached to it can be clues identifying the sense of the verb. Therefore, the above-mentioned extension will make the method applicable to disambiguating polysemous verbs. However, we need to consider which verbs to apply it to. On the one hand, disambiguating senses of most common verbs such as "get," "make," and "take" seems beyond the capability of the present method. Their correspondence between two languages is very complicated. On the other hand, verbs having domain-specific senses can be disambiguated like nouns.

Definitions of some senses of an example verb, "convert," are shown below along with typical clues identifying the senses, where OBJ and PP stand for object and prepositional phrase, respectively.

- {convert, 転換する<*TENKAN-SURU*>, 変換する<*HENKAN-SURU*>, 変える<*KAERU*>}
  ("to change or make something change from one form, substance, or state to a different one")
  [OBJ] coal, gas, electricity, program, etc.
  [PP (to)] steel, gas, code, etc.
- {convert, 両替する<*RYOUGAE-SURU*>, 換算する<*KANSAN-SURU*>}
  ("to change one type of money into another of equal value")
  [OBJ] money, dollar, yen, etc.
  [PP (into)] dollar, yen, etc.
- {convert, 転向する<*TENKOU-SURU*>, 改宗する<*KAISHUU-SURU*>, 転向させる<*TENKOU-SASERU*>, 改心させる<*KAISHIN-SASERU*>}
  ("to change or make someone change their opinion, habit, or religion")
  [PP (from)] Catholicism, Buddhism, etc.
  [PP (to)] Catholicism, Buddhism, etc.

4.7 Related work

A variety of approaches to unsupervised WSD have been proposed. Their methodologies differ greatly depending on the types of corpora and additional information they use.

A typical unsupervised WSD method is one using a monolingual corpus and a machine-readable dictionary that provides textual definitions of the senses of target words. Yarowsky's (1995) method extracts seed clues from the sense definition for each sense, and then it repeats the classification of the training examples by using seed clues and additional clues extracted from the classified examples. It achieved a precision of around 95% for 12 target words, each of which had two senses and a large number of training examples (407 to 11,968). A large number of training examples was needed because of the data sparseness problem. Other results, e.g., ones for words with more than two senses or ones for words with a small number of training examples, have not been reported. In addition, using good seed clues is essential for this method.

Karov and Edelman (1998) developed a method using a monolingual corpus and seed clues that is applicable to target words with a relatively small number of examples. The essence of their method is iterative calculation of word similarity and context similarity, through which training examples are classified. In addition, the training set is

augmented with additional examples that do not contain the target word but contain seed clues found in the sense definition. A precision of 92% was achieved for four target words, each of which had two senses and a relatively small number of training examples (27, 92, 148, and 233—excluding the additional examples). Both their method and ours are characterized by iterative calculation procedures using associations between clues, which enable them to overcome the data sparseness problem.

Schuetze (1998) developed a unique method that does not require resources other than a monolingual corpus. It clusters documents containing the target word. The resultant clusters are considered to represent respective senses of the target word, although they are not labeled as such explicitly. The average precision exceeded 80% for 10 target words, each of which had 2 senses and 1,618 to 21,374 training examples. The method, which does not take a special measure against the data sparseness problem, needs a large number of documents. It takes a huge amount of computation to cluster a large number of documents, each of which is represented with a vector having a few thousand dimensions.

Prior to the above work using monolingual corpora, the idea of WSD using parallel corpora was proposed (Brown et al. 1991b; Gale, et al. 1992b). However, it has not been pursued further. Taking the limited availability of parallel corpora into account, Dagan and Itai (1994) proposed a method using a second-language monolingual corpus and a bilingual dictionary. Strictly speaking, it is not word sense disambiguation but translated-word selection in machine translation. It produces all possible pairs of translation equivalents for a pair of syntactically related words by consulting the bilingual dictionary and selects the one with the highest co-occurrence frequency in the second-language corpus. It achieved an applicability of 68% and a precision of 91% for 103 examples in a Hebrew-to-English translation experiment. The rationale for their method is similar to ours. The difference is that it does not have a learning stage like the calculation of sense-vs.-clue correlation matrices. Therefore, it is hampered by correspondence ambiguity as well as topical coverage disparity between texts to be translated and the second-language corpus.

Kikui (1998) developed a variant of Dagan and Itai's method and applied it to term-list translation from English to Japanese. It produces all possible sets of translation equivalents, one for each term in the list, and selects then one that maximizes the average correlation among members. The correlation between words is defined as the similarity between their co-occurring word vectors. The underlying assumption is similar to ours. However, he applies it to the second language, not to the first.

Our work described here was first to demonstrate the feasibility of unsupervised

WSD using bilingual comparable corpora. The precision achieved so far is less than that of cutting-edge methods using monolingual corpora, mainly due to the uncertain correspondence between corpora of two languages. Although we have overcome the difficulty to some extent, further improvement is required. Our approach does have a particular advantage, however. It can be extended to a combined method of automatic word sense acquisition and disambiguation, as will be described in the next chapter. Schuetze's method also acquires word senses, but it does not produce definitions of the senses. In addition, our method can be applied immediately to cross-language NLP tasks. While Dagan and Itai's method is also suited to cross-language applications, it does not enable concept sharing by humans and machines, the importance of which was mentioned in Subsection 2.1.3.

4.8 Summary

We developed a method for calculating correlations between the senses of a polysemous word and the clues identifying the sense of that word based on a bilingual comparable corpus. The senses of the target polysemous word, which are input, are defined using sets of translation equivalents that represent the respective senses. The method consists of the following steps: extract word associations from the texts of each language, align the word associations by using a bilingual dictionary, and calculate the correlations between the senses and the clues, i.e., the words associated with the target word, based on the aligned word associations.

The method is characterized by its iterative calculation algorithm that uses two kinds of sets of accompanying words: the sets of translingual alignable accompanying words, each of which characterizes a word association alignment, and the sets of accompanying words regardless of translingual alignability, each of which characterizes a word association. The former resolve the ambiguity in alignment of word associations, and the latter compensate for alignment failure caused by a disparity in topical coverage between the two language texts and incomplete coverage of the bilingual dictionary. In addition, the iterative algorithm smoothes out the sparse word association data. Therefore, it is effective for target words that do not occur very frequently in the corpus. The computational load is moderate, since it deals with a relatively small sense-vs.-clue correlation matrix for each target word.

We also proposed WSD using the sense-vs.-clue correlation matrix. That is, the score of each sense of the target word is defined as the weighted sum of the correlations of the sense with clues in the context, where the weight depends on the distance between the target word and the clue, and the sense that maximizes the score is selected

for each instance of the target word. An experiment using Wall Street Journal and Nihon Keizai Shimbun corpora showed that the proposed method has promising performance: it achieved 95.8% applicability and 76.2% precision, compared to a baseline performance of 100% applicability and 62.8% precision.

Chapter 5

Clustering of Translation Equivalents Based on Similarity of Translingual Distribution Patterns

5.1 Goal and approach

While there has been a great deal of research on word sense disambiguation, there has been little on automatic word sense acquisition. Word sense acquisition has been a human activity; inventories of word senses have been constructed by lexicographers based on their best judgment. However, manually constructing an inventory of word senses is costly, the division of word senses can be arbitrary, and the word sense inventories may not match the application domains.

We address the problem of word sense acquisition as a subtask of producing a word sense association network, in which word senses are defined as sets of translation equivalents in another language. Although conventional bilingual dictionaries usually group translations according to their senses, the groupings differ by dictionary. In addition, senses specific to a domain are often missing while many senses irrelevant to the domain or unusual senses are often included. Therefore, it is best to cluster the translation equivalents of each target word based on a corpus.

We are aware of two related works on automatic word sense acquisition (Fukumoto and Tsujii 1994; Pantel and Lin 2002). Both used distributional word clustering algorithms to acquire word senses, defined as sets of synonyms, from corpora. While these algorithms seem suitable for our purpose (applying them to a set of translation equivalents of a target word results in a number of sets of synonymous translation equivalents, each of which defines one of the senses of the target word), they have insufficient cluster quality and are inapplicable to lower frequency words. Therefore, we have taken a slightly different approach; we use bilingual comparable corpora to overcome those problems.

Our proposed method, which is based on the sense-vs.-clue correlation calculation method described in Chapter 4, is outlined in Fig. 5.1. Translation equivalents of the target word are clustered hierarchically. A distinguishing feature of our method is that the similarity between (clusters of) translation equivalents is evaluated based on their

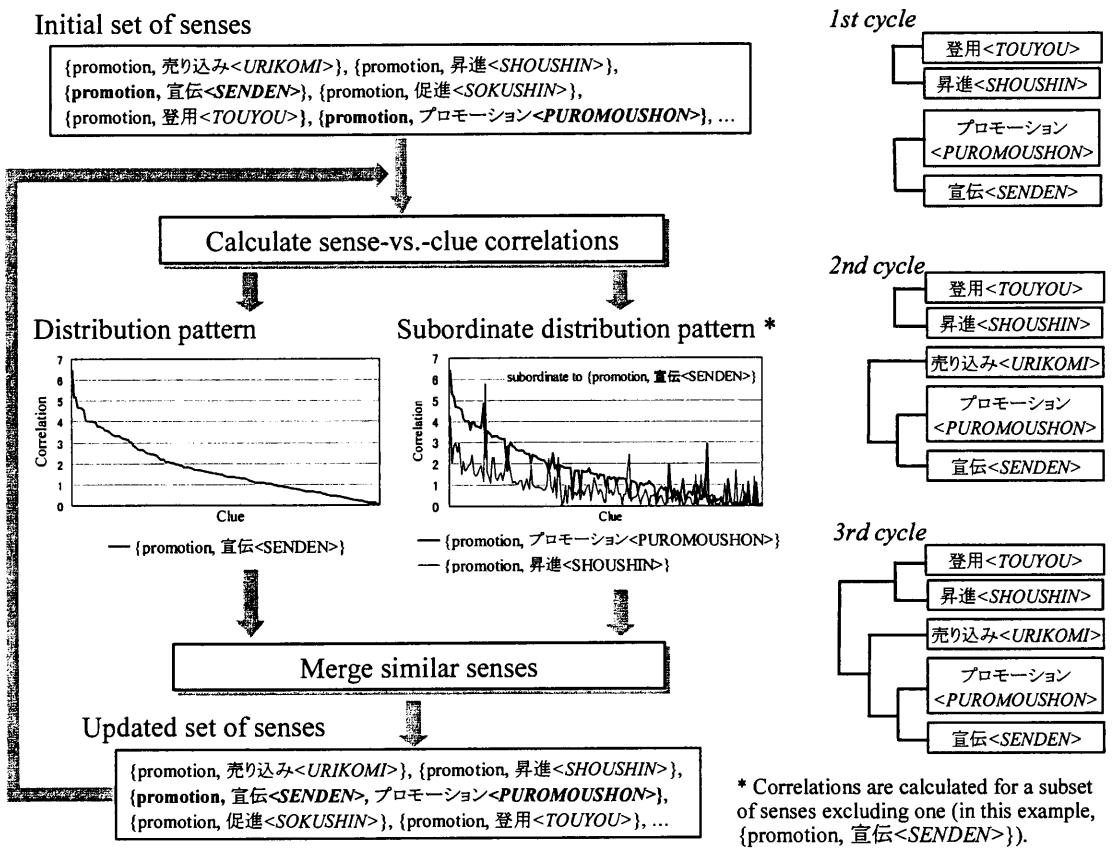


Figure 5.1 Outline of proposed method for clustering translation equivalents

translingually aligned distribution patterns. The basic idea and the algorithm will be described in detail in Sections 5.2 and 5.3, respectively.

5.2 Basic idea

5.2.1 Clustering translation equivalents of target word

Most work on automatic extraction of synonyms from text corpora rests on the idea that synonyms have similar distribution patterns (Hindle 1990; Peraira, et al. 1993; Grefenstette 1994). This idea is also useful for our task, i.e., extracting sets of synonymous translation equivalents, and we have adopted the distributional word clustering method.

We should point out that the singularity of our task makes the problems easier to solve. First, we do not have to cluster all words in a language; we only have to cluster a limited number of translation equivalents for each target word whose senses are to be acquired. As a result, the problem of computational efficiency is less serious. Note that

the amount of computation for clustering usually increases rapidly with the number of elements to be clustered (Jain, et al. 1999).

Second, even if a translation equivalent itself is polysemous, we are interested only in its sense(s) relevant to the target word. Most translation equivalents represent one and only one sense of the target word, at least in the case where the language-pair contains words with different origins, like English and Japanese. Therefore, a non-overlapping clustering algorithm, which is far simpler than overlapping clustering algorithms, is sufficient.

5.2.2 Translingual distributional word clustering

In conventional distributional word clustering, a word is characterized by a vector or weighted set consisting of words in the same language as that of the word itself. In contrast, we propose translingual distributional word clustering in which a word is characterized by a vector or weighted set consisting of words in another language. The sense-vs.-clue correlation matrix described in Chapter 4 provides the basis for doing this. The senses of a target word, as defined using its second-language translation equivalents, are characterized by the corresponding rows of a sense-vs.-clue correlation matrix, each of which is a vector or weighted sets of first-language words. That is, the second-language translation equivalents are characterized by weighted sets of first-language words.

Translingual distributional word clustering has advantages over conventional monolingual distributional word clustering for clustering translation equivalents of a target word. First, clusters are not degraded by polysemous translation equivalents. Let "race" be the target word. It has the polysemous translation equivalent "レース *<REESU>*." With monolingual distributional word clustering, "レース*<REESU>*" is characterized by a mixture of the distribution pattern for "レース*<REESU>*" representing "race" and that for "レース*<REESU>*" representing "lace," which often results in degraded clusters. In contrast, with translingual distributional word clustering, "レース *<REESU>*" is characterized by the distribution pattern for the sense of "race" that means "competition." Strictly speaking, translingual distributional word clustering clusters not words but word senses. Therefore, it is free from the problem of polysemous translation equivalents.

Second, translingual distributional word clustering can exclude from the clusters translation equivalents irrelevant to the corpus. For example, a bilingual dictionary may render "特徴*<TOKUCHOU>*" ("feature") as a translation of "race," but that sense of "race" is unusual. If it is the case in a given domain, "特徴*<TOKUCHOU>*" has low

correlation with most words associated with "race"; it can therefore be excluded from any cluster.

We should also mention the data-sparseness problem that hampers distributional word clustering. Generally speaking, the problem becomes more difficult in translingual distributional word clustering, since the sparseness of data in two languages is multiplied. However, the sense-vs.-clue correlation calculation method described in Chapter 4 overcomes this difficulty; its iterative calculation procedure smoothes out the sparse data.

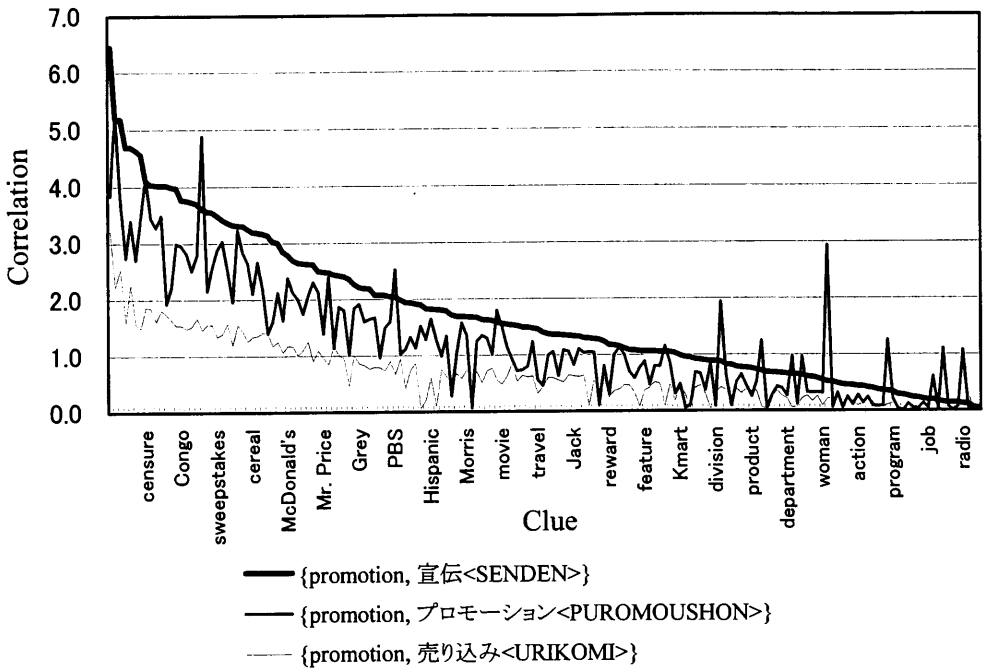
5.2.3 Similarity based on distribution pattern and subordinate distribution pattern

Naive translingual distributional word clustering based on the sense-vs.-clue correlation matrix consists of the following steps:

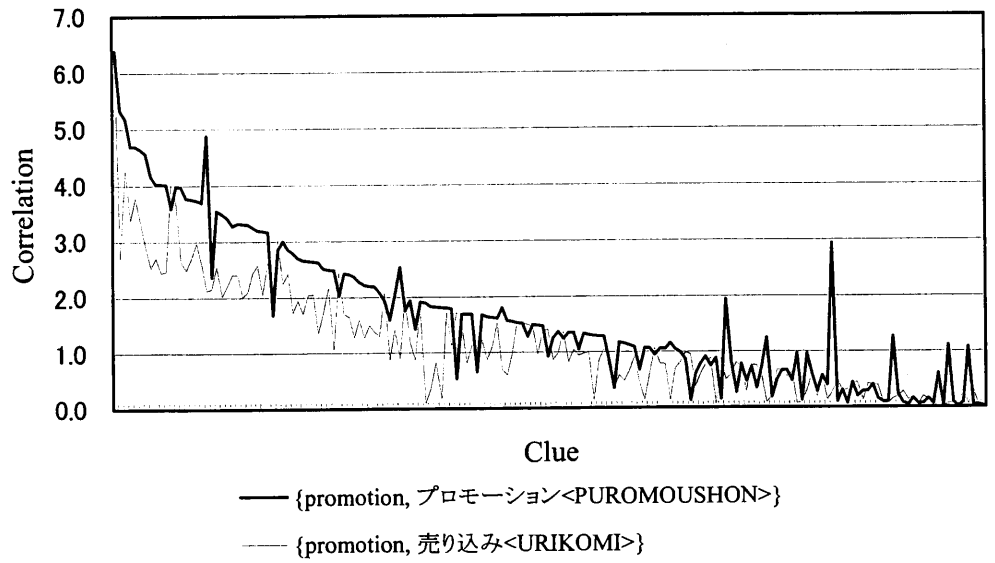
1) Define the sense of the target word by using each translation equivalent.
2) Calculate the sense-vs.-clue correlation matrix for the set of senses of the target word.
3) Calculate the similarities between senses on the basis of their distribution patterns, as shown by the sense-vs.-clue correlation matrix.
4) Cluster senses by using a hierarchical agglomerative clustering method, e.g., the group-average method.

This naive method is not effective because the algorithm for calculating the sense-vs.-clue correlation matrix presupposes a set of senses without duplicate definitions while senses are defined in duplicate in step 1). The algorithm is based on the "one sense per collocation" hypothesis, so it results in each clue having a high correlation with one and only one sense. Two or more senses cannot be highly correlated with the same clue, even when they are actually the same sense. Therefore, senses defined with synonymous translation equivalents do not necessarily have very high similarity.

Figure 5.2(a) shows distribution patterns for {promotion, 宣伝 <*SENDEN*>}, {promotion, プロモーション <*PUROMOUSHON*>}, and {promotion, 売り込み <*URIKOMI*>}, all of which define the "sales activity" sense of "promotion." The horizontal axis represents a set of clues, sorted in descending order of correlation with {promotion, 宣伝<*SENDEN*>}, and the vertical axis represents the correlation between senses and clues. Most clues identifying the "sales activity" sense have the highest correlation with {promotion, 宣伝<*SENDEN*>} and have relatively low correlation with {promotion, プロモーション<*PUROMOUSHON*>} and {promotion, 売り込み<*URIKOMI*>}, because "宣伝<*SENDEN*>" is the most dominant translation equivalent of "promotion"



(a) Distribution patterns



(b) Distribution patterns subordinate to {promotion, 宣伝<*SENDEN*>}

Figure 5.2 Distribution pattern vs. subordinate distribution pattern

representing that sense. As a result, these three senses do not have very high similarity.

To resolve this problem, we need to calculate the sense-vs.-clue correlation matrix

not only for the full set of senses but also for the sets of senses excluding one of these senses. Excluding a sense defined with the most dominant translation equivalent allows most clues identifying the sense to have the highest correlation with another sense defined with the second most dominant translation equivalent.

Figure 5.2(b) shows distribution patterns for {promotion, プロモーション <*PUROMOUSHON*>} and {promotion, 売り込み<*URIKOMI*>} when the sense-vs.-clue correlation matrix is calculated for the set of senses excluding {promotion, 宣伝 <*SENDEN*>}. In this case, most clues identifying the "sales activity" sense have the highest correlation with {promotion, プロモーション<*PUROMOUSHON*>}, because "プロモーション<*PUROMOUSHON*>" is the second most dominant translation equivalent of "promotion" representing that sense. Note that the distribution pattern for {promotion, プロモーション<*PUROMOUSHON*>} in Fig. 5.2(b) is more similar to that for {promotion, 宣伝 <*SENDEN*>} in Fig. 5.2(a) than that for {promotion, プロモーション <*PUROMOUSHON*>} in Fig. 5.2(a).

The distribution pattern for sense S_2 shown by the sense-vs.-clue correlation matrix for the set of senses excluding sense S_1 is called the distribution pattern for S_2 subordinate to S_1. The distribution pattern for sense S_2 shown by the sense-vs.-clue correlation matrix for the full set of senses is called simply the distribution pattern for S_2. The similarity of S_2 to S_1 is defined as the similarity of the distribution pattern for S_2 subordinate to S_1 to the distribution pattern for S_1.

Calculating the sense-vs.-clue correlation matrix for a set of senses excluding one sense is, of course, insufficient since three or more translation equivalents may represent the same sense of the target word. We need to merge similar senses into one sense and then recalculate the sense-vs.-clue correlation matrices both for the full set of senses and for the set of senses excluding one of these senses. Repeating these steps enables corpus-relevant but less dominant translation equivalents to move up, while corpus-irrelevant ones do not. Thus, corpus-relevant translation equivalents of the target word are hierarchically clustered.

5.3 Algorithm

5.3.1 Outline

As shown in Fig. 5.3, our proposed method repeats the following three steps:

1) Calculate sense-vs.-clue correlation matrices both for the full set of senses and for sets of senses excluding one of these senses.
2) Calculate similarities between senses based on distribution patterns shown by the sense-vs.-clue correlation matrix for the full set of senses and subordinate

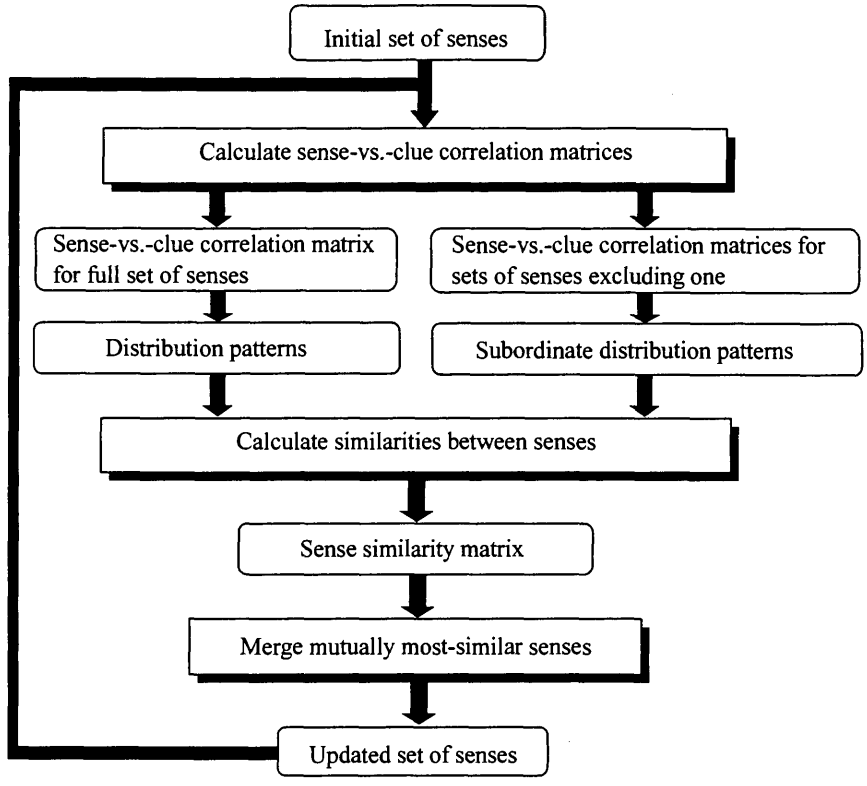


Figure 5.3 Flow of translation equivalent clustering

distribution patterns shown by the sense-vs.-clue correlation matrices for sets of senses excluding one.

3) Merge every pair of mutually most-similar senses into one sense.

The initial set of senses is given as $\Sigma(x)=\{\{x, y_1\}, \{x, y_2\}, \ldots, \{x, y_N\}\}$, where x is a target word in the first language, and $y_1, y_2, \ldots$, and y_N are translation equivalents of x in the second language. Translation equivalents that occur less frequently in the second-language corpus can be excluded from the initial set to shorten the processing time. These steps are described in detail in the following subsections.

5.3.2 Calculation of sense-vs.-clue correlation matrices

A sense-vs.-clue correlation matrix is calculated for the full set of senses. The resulting correlation matrix is denoted as C. That is, $C(i, j)$ is the correlation between the i-th sense, $S(x, i)$, of target word x and its j-th clue, $x'(j)$.

Then, the set of relevant senses, $\Sigma_A(x)$, is determined. A sense is regarded relevant to the corpus if and only if the ratio of clues with which it has the highest correlation exceeds a predetermined threshold, θ. (In the experiment described in Section 5.4, θ was

set to 0.05). That is,

$$\Sigma_A(x) = \{S(x,i) \mid R(S(x,i)) > \theta\},$$

where $R(S(x, i))$ denotes the ratio of clues having the highest correlation with $S(x, i)$, i.e.,

$$R(S(x,i)) = \frac{\left|\{x'(j) \mid C(i,j) = \max_k C(k,j)\}\right|}{\left|\{x'(j)\}\right|}.$$

If only one sense is relevant, the clustering procedure terminates. Otherwise, a sense-vs.-clue correlation matrix is calculated for the set of senses excluding each of the relevant senses. The correlation matrix calculated for the set of senses excluding the k-th sense is denoted as C_{-k}. That is, $C_{-k}(i, j)$ $(i \neq k)$ is the correlation between the i-th sense and the j-th clue, calculated excluding the k-th sense. $C_{-k}(k, j)$ $(j=1, 2,...)$ are set to zero. This redundant k-th row is included to maintain the same correspondence between rows and senses as in C.

5.3.3 Calculation of sense similarity matrix

The similarity of the i-th sense to the j-th sense, $Sim(S(x, i), S(x, j))$, is defined as the similarity of the distribution pattern for $S(x, i)$ subordinate to $S(x, j)$ to the distribution pattern for $S(x, j)$. Note that this similarity is asymmetric and reflects which sense is more dominant in the corpus. It is probable that $Sim(S(x, i), S(x, j))$ is large while $Sim(S(x, j), S(x, i))$ is not when $S(x, j)$ is more dominant than $S(x, i)$.

According to the sense-vs.-clue correlation matrix, each sense is characterized using a weighted set of clues. Therefore, we used the weighted Jaccard coefficient as the similarity measure. That is,

$$Sim(S(x,i),S(x,j)) = \frac{\sum_k \min\{C_{-j}(i,k), C(j,k)\}}{\sum_k \max\{C_{-j}(i,k), C(j,k)\}} \quad \text{when } S(x,j) \in \Sigma_A(x).$$

$$Sim(S(x,i),S(x,j)) = 0 \quad \text{otherwise.}$$

It should be noted that a sense is characterized by different weighted sets of clues depending on which sense the similarity is calculated. Note also that only the similarities of senses to relevant senses are of concern.

5.3.4 Merging similar senses

The set of senses is updated by merging every pair of senses with mutually highest similarity into one. That is,

$$\Sigma(x) \leftarrow \Sigma(x) - \{S(x, i), S(x,j)\} + \{S(x, i) \cup S(x,j)\}$$

if $Sim(S(x,i),S(x,j)) = \max_{j'}\{\max\{Sim(S(x,i),S(x,j')),Sim(S(x,j'),S(x,i))\}\} > \sigma$,

$Sim(S(x,i),S(x,j)) = \max_{i'}\{\max\{Sim(S(x,i'),S(x,j)),Sim(S(x,j),S(x,i'))\}\} > \sigma$,

and $Sim(S(x,i),S(x,j)) > \sigma$.

The σ is a predetermined threshold for similarity; it is introduced to avoid noisy pairs of senses being merged. In the experiment described in Section 5.4, it was set to 0.25.

If at least one pair of senses is merged, the whole procedure, i.e., the calculation of sense-vs.-clue correlation matrices through the merger of similar senses, is repeated for the updated set of senses. Otherwise, the clustering procedure terminates.

Agglomerative clustering methods usually suffer from the problem of when to terminate merging. In our method, the similarity of senses merged into one does not necessarily decrease monotonically, which makes the problem more difficult. At present, we are forced to output a dendrogram representing the history of mergers and leave the final decision to a person. The dendrogram consists of the translation equivalents defining the relevant sense(s) after the final cycle.

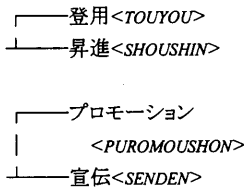
5.3.5 Illustrative example of clustering

Figure 5.4 illustrates how the clustering proceeds for the target word "promotion." Figures 5.4(a-1), (b-1), and (c-1) show the sense similarity matrices in the first, second, and third cycles, and Figs. 5.4(a-2), (b-2), and (c-2) show the clustering results of the first, second, and third cycles.

The initial set of senses consists of 12 senses, which are defined using the Japanese translation equivalents of the target word. Four of the 12 senses are relevant, and calculating the sense similarity matrix (Fig. 5.4(a-1)) results in two pairs of mutually most-similar senses, i.e., {promotion, 宣伝<*SENDEN*>} and {promotion, プロモーション<*PUROMOUSHON*>}, and {promotion, 昇進<*SHOUSHIN*>} and {promotion, 登用<*TOUYOU*>}. Merging these pairs (Fig. 5.4(a-2)) results in ten senses, of which two are relevant.

Sense \ Relevant sense	宣伝 <SENDEN> [6]	昇進 <SHOUSHIN> [7]	登用 <TOUYOU> [8]	プロモーション <PUROMOUSHON> [10]
就任<SHUUNIN> [1]	0.371	0.679	0.708	0.390
促進<SOKUSHIIN> [2]	0.324	0.559	0.502	0.338
昇格<SHOUKAKU> [3]	0.356	0.739	0.686	0.377
振興<SHINKOU> [4]	0.392	0.682	0.616	0.398
売り込み<URIKOMI> [5]	0.693	0.661	0.763	0.557
宣伝<SENDEN> [6]	-	0.370	0.422	0.704
昇進<SHOUSHIN> [7]	0.457	-	0.747	0.429
登用<TOUYOU> [8]	0.562	***0.772***	-	0.521
奨励<SHOUREI> [9]	0.476	0.724	0.621	0.408
プロモーション <PUROMOUSHON> [10]	***0.903***	0.420	0.507	-
増進<ZOUSHIN> [11]	0.430	0.699	0.664	0.418
助長<JOCHOU> [12]	0.285	0.514	0.453	0.301

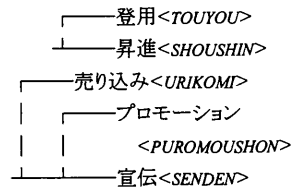
(a-1) Sense similarity matrix in 1st cycle



(a-2) Results of 1st cycle

Sense \ Relevant sense	宣伝<SENDEN> [6], プロモーション <PUROMOUSHON> [10]	昇進<SHOUSHIN> [7], 登用<TOUYOU> [8]
就任<SHUUNIN> [1]	0.530	0.562
促進<SOKUSHIIN> [2]	0.492	0.358
昇格<SHOUKAKU> [3]	0.482	0.652
振興<SHINKOU> [4]	0.574	0.500
売り込み<URIKOMI> [5]	***0.932***	0.613
宣伝<SENDEN> [6], プロモーション<PUROMOUSHON> [10]	-	0.427
昇進<SHOUSHIN> [7], 登用<TOUYOU> [8]	0.763	-
奨励<SHOUREI> [9]	0.683	0.493
増進<ZOUSHIN> [11]	0.717	0.498
助長<JOCHOU> [12]	0.416	0.330

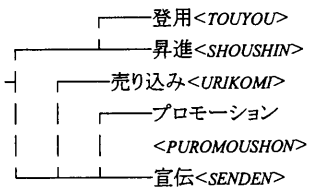
(b-1) Sense similarity matrix in 2nd cycle



(b-2) Results of 2nd cycle

Sense \ Relevant sense	宣伝<SENDEN> [6], プロモーション <PUROMOUSHON> [10], 売り込み<URIKOMI> [5]	昇進<SHOUSHIN> [7], 登用<TOUYOU> [8]
就任<SHUUNIN> [1]	0.583	0.558
促進<SOKUSHIIN> [2]	0.612	0.354
昇格<SHOUKAKU> [3]	0.550	0.641
振興<SHINKOU> [4]	0.668	0.490
宣伝<SENDEN> [6], プロモーション<PUROMOUSHON> [10], 売り込み<URIKOMI> [5]	-	0.421
昇進<SHOUSHIN> [7], 登用<TOUYOU> [8]	***0.885***	-
奨励<SHOUREI> [9]	0.832	0.484
増進<ZOUSHIN> [11]	0.865	0.484
助長<JOCHOU> [12]	0.471	0.324

(c-1) Sense similarity matrix in 3rd cycle



(c-2) Results of 3rd cycle

[Note 1] Definitions of senses are shown excluding target word "promotion" due to limited space.
[Note 2] For each Japanese translation equivalent, an English equivalent other than the target word is shown:
1) taking up a post, 2) acceleration, 3) raising, 4) furtherance, 5) sale, 6) advertisement, 7) advancement, 8) elevation, 9) encouragement, 10) advertising campaign, 11) advance, 12) furtherance.

Figure 5.4 Clustering translation equivalents of "promotion"

In the second cycle, the sense similarity matrix is calculated for the updated set of senses (Fig. 5.4(b-1)). As a result, a pair of mutually most-similar senses, {promotion, 宣伝<*SENDEN*>, プロモーション<*PUROMOUSHON*>} and {promotion, 売り込み<*URIKOMI*>}, is obtained. Merging this pair (Fig. 5.4(b-2)) results in nine senses, of which two are relevant. Note that the similarity of {promotion, 売り込み<*URIKOMI*>} to {promotion, 宣伝<*SENDEN*>, プロモーション<*PUROMOUSHON*>} (0.932) is larger than that of {promotion, プロモーション<*PUROMOUSHON*>} to {promotion, 宣伝<*SENDEN*>} (0.903) in the first cycle. The similarity of senses to be merged does not necessarily decrease monotonically, unlike in conventional hierarchical agglomerative clustering.

Likewise, in the third cycle, the sense similarity matrix is calculated for the updated set of senses (Fig. 5.4(c-1)). As a result, a pair of mutually most-similar senses, {promotion, 宣伝<*SENDEN*>, プロモーション<*PUROMOUSHON*>, 売り込み<*URIKOMI*>} and {promotion, 昇進<*SHOUSHIN*>, 登用<*TOUYOU*>}, is obtained. Merging this pair (Fig. 5.4(c-2)) results in eight senses, of which only one is relevant. Therefore, the clustering procedure is terminated, and a dendrogram consisting of "宣伝<*SENDEN*>," "プロモーション<*PUROMOUSHON*>," 売り込み<*URIKOMI*>," "昇進<*SHOUSHIN*>," and "登用<*TOUYOU*>," which define the relevant sense, is output.

5.3.6 Variations

We have alternatives to the definition of similarity as well as the criterion for selecting senses to be merged.

(1) Asymmetric similarity versus symmetric similarity

The similarity between the i-th sense, $S(x, i)$, and the j-th sense, $S(x, j)$, can also be defined as the similarity between the distribution pattern for $S(x, i)$ subordinate to $S(x, j)$ and that for $S(x, j)$ subordinate to $S(x, i)$. That is,

$$Sim(S(x,i),S(x,j)) = \frac{\sum_k \min\{C_{-j}(i,k),C_{-i}(j,k)\}}{\sum_k \max\{C_{-j}(i,k),C_{-i}(j,k)\}} \quad \text{when } S(x,i)\in\Sigma_A(x) \text{ or } S(x,j)\in\Sigma_A(x).$$

$$Sim(S(x,i),S(x,j)) = 0 \quad \text{otherwise.}$$

This symmetric similarity seems natural as it indicates the substitutionability of senses for each other. It is probably better than the asymmetric similarity described in Subsection 5.3.3 when the dominance relation is not manifest among translation equivalents representing the same sense. Its deficiency is that a pair of distinctive senses, neither of which has duplicate definitions, can have high similarity.

(2) Weighted Jaccard coefficient versus Jaccard coefficient

According to the "one sense per collocation" hypothesis, it may be better to assign each clue to the sense with the highest correlation. That is, the sense-vs.-clue correlation matrix for the full set of senses, $C(i,j)$, is converted into a binary matrix:

$$C(i,j) \leftarrow 1 \quad \text{if} \quad C(i,j) = \max_{i'} C(i',j).$$

$$C(i,j) \leftarrow 0 \quad \text{otherwise}.$$

Likewise, the sense-vs.-clue correlation matrix for the set of senses excluding the k-th sense, $C_{-k}(i,j)$, is converted into a binary matrix:

$$C_{-k}(i,j) \leftarrow 1 \quad \text{if} \quad C_{-k}(i,j) = \max_{i'} C_{-k}(i',j).$$

$$C_{-k}(i,j) \leftarrow 0 \quad \text{otherwise}.$$

Using these binary matrices, we can define the similarity between senses by using the Jaccard coefficient instead of the weighted Jaccard coefficient.

(3) Mutually most-similar pairs versus the most similar pair

Merging senses can be restricted to one pair per cycle, i.e., the most similar pair. That is,

$$\Sigma(x) \leftarrow \Sigma(x) - \{S(x,i), S(x,j)\} + \{S(x,i) \cup S(x,j)\}$$

$$\text{if} \quad Sim(S(x,i),S(x,j)) = \max_{i',j'} Sim(S(x,i'),S(x,j')).$$

Combining these alternatives results in eight variations, including the original. These and two additional variations are compared in the following section. The additional variations are included to confirm the effectiveness of the subordinate distribution patterns. They use a primitive similarity defined by not using the subordinate distribution patterns. The primitive similarity between the i-th sense, $S(x,i)$, and the j-th sense, $S(x,j)$, is defined as the similarity between the distribution pattern for $S(x,i)$ and that for $S(x,j)$:

$$Sim(S(x,i),S(x,j)) = \frac{\sum_k \min\{C(i,k),C(j,k)\}}{\sum_k \max\{C(i,k),C(j,k)\}} \quad \text{when } S(x,i) \in \Sigma_A(x) \text{ or } S(x,j) \in \Sigma_A(x).$$

$$Sim(S(x,i),S(x,j)) = 0 \quad \text{otherwise}.$$

Note that the primitive similarity is incompatible with converting the sense-vs.-clue correlation matrix into a binary matrix, so it cannot be combined with the (binary) Jaccard coefficient. One of the additional variations merges mutually most-similar pairs, and the other merges the most similar pair.

5.4 Experimental evaluation

5.4.1 Method and materials

We evaluated our method experimentally using the same comparable corpus and bilingual dictionary described in Subsection 4.5.1, i.e., a combination of a 189-MB Wall Street Journal corpus and a 275-MB Nihon Keizai Shimbun corpus, and the English-Japanese noun dictionary resulting from the merger of the EDR English-to-Japanese and Japanese-to-English dictionaries. Extraction of word associations from the corpus of each language was done using the same settings as in Subsection 4.5.1. The sense-vs.-clue correlation matrices were calculated using Method II with $\alpha = 5$ and $N = 6$.

Evaluating the performance of word sense acquisition methods is not a trivial task. First, we do not have a gold-standard sense inventory. Even if we had one, we would have difficulty mapping the acquired senses onto the senses in it. Note that any sense can be defined differently from the definition given by the standard sense inventory. Second, there is no way to establish the complete set of senses appearing in a large corpus. Therefore, we evaluated our method using a limited number of target words, i.e., the 60 English polysemous words we selected for the WSD experiment described in Section 4.5. The collection of sense definitions used in that experiment was used as the standard sense inventory. The collection of manually sense-tagged instances of the target words used in the WSD experiment was also used to estimate the ratios of senses of each target word in the corpus. We regarded senses with ratios not less than a certain threshold as those to be acquired.

5.4.2 Evaluative measures

We defined two evaluative measures: recall of senses and accuracy of sense definitions.

(1) Recall of senses

The recall of senses is the proportion of senses with ratios not less than a threshold that are successfully acquired. It varies with change in the threshold. We judged that a sense was acquired when the output dendrogram of translation equivalents included at least one translation equivalent defining it.

Table 5.1(a) illustrates the measurement of the recall of senses for the example target word "promotion." Among the three senses of "promotion" defined in the standard sense inventory, Sense 1, i.e., {promotion, 宣伝<*SENDEN*>, 売り込み<*URIKOMI*>, 販売促進<*HANBAI-SOKUSHIN*>, プロモーション<*PUROMOUSHON*>}, was judged to be acquired, since the output dendrogram included three out of the four translation equivalents defining it. Sense 2, i.e., {promotion, 昇格<*SHOUKAKU*>, 昇進<*SHOUSHIN*>, 昇任<*SHOUNIN*>, 就任<*SHUUNIN*>, 登用<*TOUYOU*>, 進級

<*SHINKYUU*>}, was also judged to be acquired, since the output dendrogram included two out of the six translation equivalents defining it. In contrast, Sense 3, i.e., {promotion, 奨励<*SHOUREI*>, 振興<*SHINKOU*>, 促進<*SOKUSHIN*>, 増進<*ZOUSHIN*>, 助長<*JOCHOU*>}, was judged not to be acquired, since the output dendrogram included none of the five translation equivalents defining it. Given that the ratio of the third sense in the corpus was 0.03, the results seem reasonable. The recall of senses was 1.00 for a sense-ratio threshold larger than 0.03, while it was 0.67 for a sense-ratio threshold less than and equal to 0.03.

(2) Accuracy of sense definitions

To evaluate the accuracy of sense definitions while avoiding mapping acquired senses onto those defined in the standard sense inventory, we regarded a set of sense definitions as a set consisting of pairs of translation equivalents that define the same sense. Let T_S be a set of pairs of translation equivalents defining the same sense in the standard sense inventory. Likewise, let $T(k)$ be a set of pairs of translation equivalents defining the same relevant sense in the k-th cycle of the clustering procedure. Furthermore, let U be a set of pairs of translation equivalents that are included in the output dendrogram. Recall, precision, and the F-measure of pairs of translation equivalents defining the same sense in the k-th cycle are defined, respectively, as

$$R(k) = \frac{|T_S \cap T(k)|}{|T_S \cap U|},$$

$$P(k) = \frac{|T_S \cap T(k)|}{|T(k)|}, \text{ and}$$

$$F(k) = \frac{2 \cdot R(k) \cdot P(k)}{R(k) + P(k)}.$$

Note that the intersection of T_S and U is used as the denominator for the recall to exclude corpus-irrelevant translation equivalents.

The F-measure indicates how well the set of relevant senses in each cycle coincides with the set of sense definitions in the standard sense inventory. Although the present method cannot stop the clustering procedure at the end of the optimum cycle, a person can identify the set of appropriate senses from the output dendrogram at a glance. We thus defined the accuracy of sense definitions as the maximum F-measure in all cycles.

Table 5.1(b) illustrates the measurement of the accuracy of sense definitions for the example target word "promotion." As shown in the table, the F-measures in the first, second, and third cycles were 0.67, 1.00, and 0.57, respectively. As a result, the accuracy of sense definitions was 1.00. Note that terminating the clustering procedure at the end of the second cycle would result in perfect results.

Table 5.1 Evaluation of word sense acquisition for "promotion"

(a) Recall of senses

Standard sense inventory / Output dendrogram	Sense 1 "an activity intended to help sell a product"				Sense 2 "advancement in rank or position"						Sense 3 "action to help something develop or succeed"				
	宣伝<*SENDEN*>	売り込み<*URIKOMI*>	販売促進<*HANABI-SOKUSHIN*>	プロモーション<*PUROMOUSHON*>	昇格<*SHOUKAKU*>	昇進<*SHOUSHIN*>	昇任<*SHOUNIN*>	就任<*SHUUNIN*>	登用<*TOUYOU*>	進級<*SHINKYUU*>	奨励<*SHOUREI*>	振興<*SHINKOU*>	促進<*SIKUSHIIN*>	増進<*ZOUSHIN*>	助長<*JOCHOU*>
登用<*TOUYOU*>									√						
昇進<*SHOUSHIN*>						√									
売り込み<*URIKOMI*>		√													
プロモーション <*PUROMOUSHON*>				√											
宣伝<*SENDEN*>	√														
Judgment	Acquired				Acquired						Not acquired				
Sense ratio in corpus	0.73				0.24						0.03				
Recall of senses	1.00 for sense-ratio threshold larger than 0.03 0.67 for sense-ratio threshold less than and eqaul to 0.03														

(b) Accuracy of sense definitions

Pair of translation equivalents defining the same sense	$TS \cap U$	$T(1)$	$T(2)$	$T(3)$
(宣伝<*SENDEN*>, プロモーション<*PUROMOUSHON*>)	√	√	√	√
(昇進<*SHOUSHIN*>, 登用<*TOUYOU*>)	√	√	√	√
(宣伝<*SENDEN*>, 売り込み<*URIKOMI*>)	√		√	√
(売り込み<*URIKOMI*>, プロモーション<*PUROMOUSHON*>)	√		√	√
(宣伝<*SENDEN*>, 昇進<*SHOUSHIN*>)				√
(宣伝<*SENDEN*>, 登用<*TOUYOU*>)				√
(プロモーション<*PUROMOUSHON*>, 昇進<*SHOUSHIN*>)				√
(プロモーション<*PUROMOUSHON*>, 登用<*TOUYOU*>)				√
(売り込み<*URIKOMI*>, 昇進<*SHOUSHIN*>)				√
(売り込み<*URIKOMI*>, 登用<*TOUYOU*>)				√
Recall: $R(k) = \| TS \cap T(k) \| / \| TS \cap U \|$		0.50	1.00	1.00
Precision: $P(k) = \| TS \cap T(k) \| / \| T(k) \|$		1.00	1.00	0.40
F-measure: $F(k) = 2 \cdot R(k) \cdot P(k) / (R(k) + P(k))$		0.67	1.00	0.57
Accuracy of sense definitions: $\max_k F(k)$		1.00		

5.4.3 Comparison between variations

To simplify the evaluation procedure, we clustered translation equivalents that were

used to define the senses of each target word in the standard sense inventory, rather than clustering translation equivalents rendered by the EDR bilingual dictionary. The translation equivalents to be clustered were restricted to those that occurred ten or more times in the training corpus. When 20 or more translation equivalents occurred 10 or more times, the 20 most frequently occurring ones were selected.

Figure 5.5(a) shows the recall of senses for the ten variations of our method. It was measured for a total of 201 senses of the 60 target words. Note that the scale on the x axis is not linear, reflecting the distribution of the sense ratios. Figure 5.5(b) shows the accuracy of sense definitions for the ten variations. It was measured for each target word and averaged over the 60 target words.

We compared the results to determine which variation is best. First, the difference between the asymmetric similarity (A) and the symmetric similarity (S) was small compared to that between them and the primitive similarity (P). We examined clusters produced by the method using the primitive similarity, which featured high recall of senses and low accuracy of sense definitions. We had great difficulty in recognizing senses from those clusters, so we rejected the use of primitive similarity. It was difficult to determine whether the asymmetric or symmetric similarity was better. The asymmetric similarity had a higher recall of senses when combined with the weighted Jaccard coefficient (WJ) and the merger of mutually most-similar pairs (MM); however, its accuracy of sense definitions was relatively low. In contrast, the symmetric similarity had lower recall of senses and higher accuracy of sense definitions.

Second, the weighted Jaccard coefficient (WJ) was generally better than the Jaccard coefficient (J). When they were combined with the same alternatives, the weighted Jaccard coefficient was almost always better than the Jaccard coefficient.

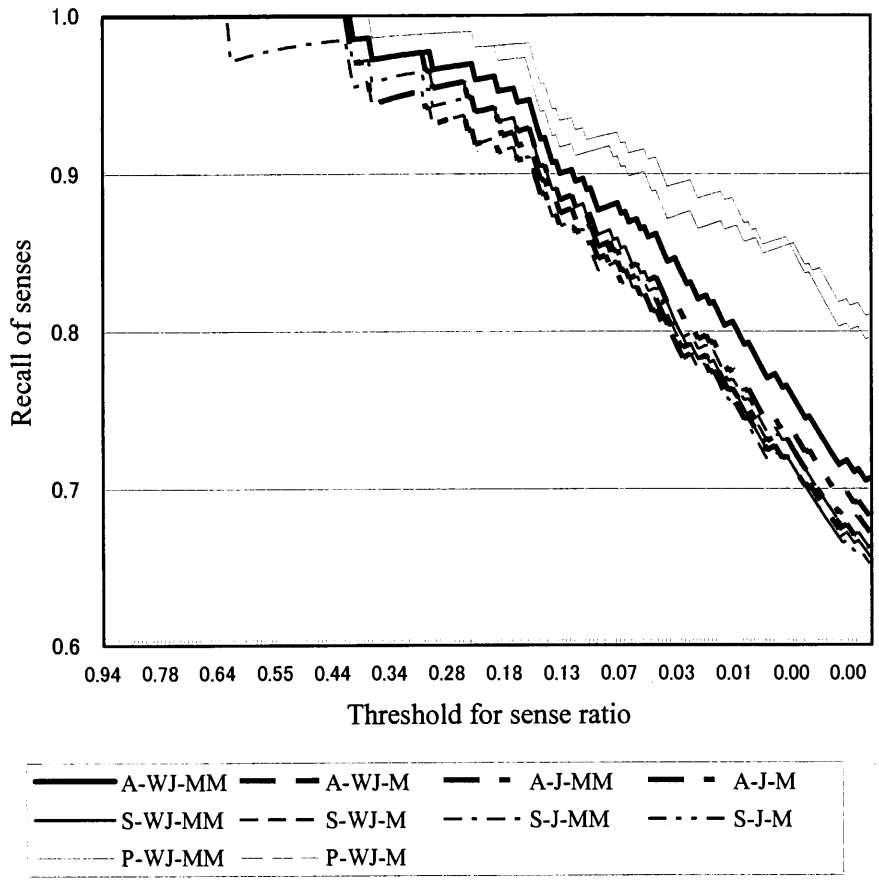
Third, there was little difference between the merger of mutually most-similar pairs (MM) and the merger of the most similar pair (M). Taking computational efficiency into account, we preferred the merger of mutually most-similar pairs because it produces a hierarchy of senses in fewer cycles.

Our choice was thus narrowed down to two variations: A-WJ-MM and S-WJ-MM. A-WJ-MM, which is the original, had a higher recall of senses:

95.9% for senses with ratios not less than 25%,

87.1% for senses with ratios not less than 5%, and

77.5% for senses with ratios not less than 1%.



(a) Recall of senses

Variation			Accuracy
A (Asymmetric similarity)	WJ (Weighted Jaccard coefficient)	MM (Mutually most-similar pairs)	0.766
		M (Most-similar pair)	0.776
	J (Jaccard coefficient)	MM	0.761
		M	0.773
S (Symmetric similarity)	WJ	MM	0.805
		M	0.798
	J	MM	0.793
		M	0.780
P (Primitive similarity)	WJ	MM	0.636
		M	0.621

(b) Accuracy of sense definitions

Figure 5.5 Recall of senses and accuracy of sense definitions for ten variations of proposed translation equivalent clustering method

The accuracy of sense definitions was 76.6%. In contrast, S-WJ-MM had higher accuracy of sense definitions, i.e., 80.5%. Its recall of senses was

93.9% for senses with ratios not less than 25%,

84.3% for senses with ratios not less than 5%, and

72.8% for senses with ratios not less than 1%.

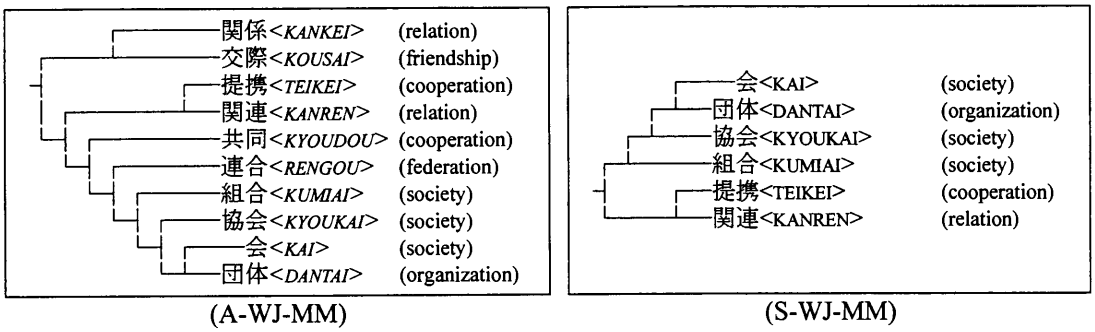
Figure 5.6 shows the clustering results for four target words using A-WJ-MM and S-WJ-MM, illustrating the contrasting strengths and weaknesses. A-WJ-MM tended to produce dendrograms consisting of a larger number of translation equivalents than S-WJ-MM. While they were sometimes noisy, they showed more senses of the target word. In contrast, S-WJ-MM tended to produce dendrograms showing fewer senses. For example, the dendrogram produced for "bar" by A-WJ-MM consisted of not only translation equivalents representing the senses of "bar" as "place to drink in" and "group of lawyers," but also "格子<*KOUSHI*>" and "障害<*SHOUGAI*>," representing the senses of "bar" as "piece of metal or wood" and "obstacle." The dendrogram produced for "bar" by S-WJ-MM consisted of only translation equivalents representing the senses of "bar" as "place to drink in" and "group of lawyers."

5.4.4 Detailed analysis of example results

While these results show that our method has a great deal of promise, they also revealed some deficiencies. The first is that it performs poorly for non-topical, generic senses. For example, despite its relatively high ratio in the corpus (25%), the sense of "bar" as "a piece of solid material" could not be acquired (See Fig. 5.6(b)). Although it was acquired, the sense of "measure" as "an action taken to gain a certain end," whose estimated ratio in the corpus was 38%, was less prominent than the other senses: "a system or instrument for calculating amount, size, weight, etc." and "a law suggested in Parliament" (See Fig. 5.6(d)). These generic senses had fewer clues identifying them than the topical senses, so they were difficult to acquire.

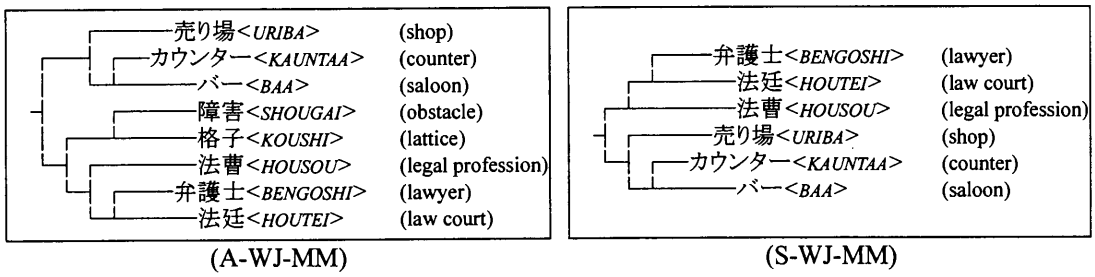
The second deficiency is that our method also performs poorly for low-frequency senses. For example, it failed to acquire the sense of "promotion" as "action to help something develop or succeed" (See Fig. 5.4), which had an estimated ratio in the corpus of only 3% (those of the senses "an activity intended to help sell a product" and "advancement in rank or position" were 73% and 24%, respectively). Using a larger corpus would improve the performance for the low-frequency senses.

The third deficiency lies in the crucial role of the bilingual dictionary. It is obvious that a sense cannot be acquired if the translation equivalents representing it are not in the dictionary. An exhaustive bilingual dictionary is therefore required. While from this point of view, the EDR bilingual dictionary worked fairly well, it was inadequate for some target words.



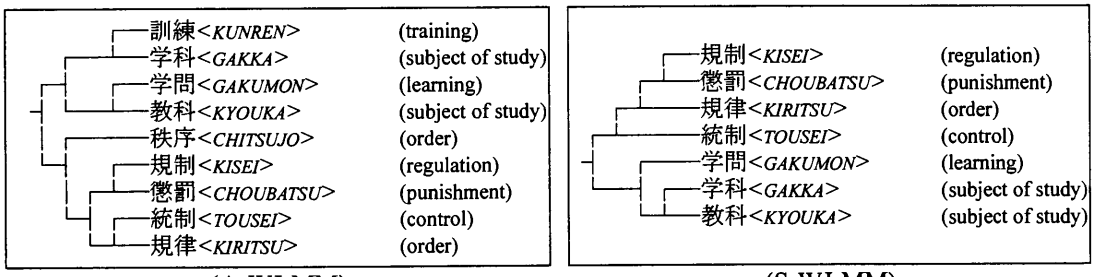
(A-WJ-MM) (S-WJ-MM)

(a) "association"



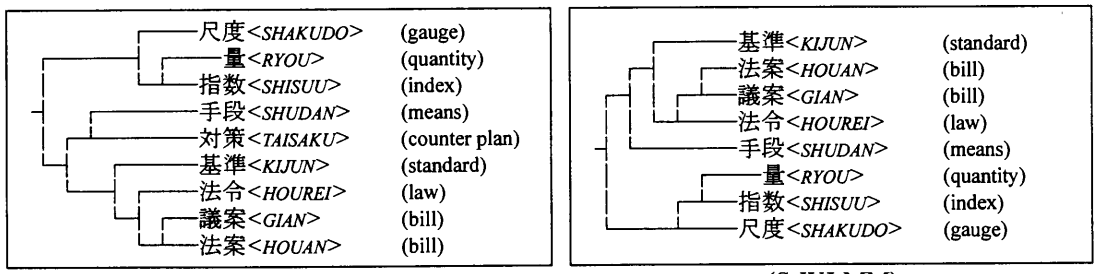
(A-WJ-MM) (S-WJ-MM)

(b) "bar"



(A-WJ-MM) (S-WJ-MM)

(c) "discipline"



(A-WJ-MM) (S-WJ-MM)

(d) "measure"

[Note] For each Japanese-translation equivalent, an English equivalent other than the target word is shown to the right.

Figure 5.6 Clustering results for four target words

Figure 5.7 shows example pairs of dendrograms our method (A-WJ-MM) produced for the same target word from different sets of translation equivalents: one consisting of translation equivalents rendered by the EDR bilingual dictionary, and the other consisting of translation equivalents rendered by two everyday English-to-Japanese dictionaries, i.e., Kenkyusha's English-Japanese Dictionary for the General Reader (KENKYUSHA 1984) and Kenkyusha's New Collegiate English-Japanese Dictionary (KENKYUSHA 1985). The entry for "coach" in the EDR bilingual dictionary did not include "エコノミークラス<*EKONOMII-KURASU*>," which represents a modern sense of "coach" ("economy class"), while it included an out-of-date translation equivalent "箱<*HAKO*>" ("railway carriage"). This and the erroneous merger of "バス<*BASU*>" and "指導<*SHIDOU*>" resulted in a less understandable dendrogram. In contrast, the alternative dictionaries, which rendered "エコノミークラス<*EKONOMII-KURASU*>" as a translation of "coach," resulted in an appropriate dendrogram. Furthermore, the entry for "wing" in the EDR bilingual dictionary did not include "派<*HA*>" and "党派<*TOUHA*>," which represent its sense as "group within an organization." As a result, despite its high ratio in the Wall Street Journal corpus (44%), this sense could not be acquired. In contrast, the alternative dictionaries, which did render "派<*HA*>" and "党派<*TOUHA*>" as translations of "wing," enabled the sense to be acquired.

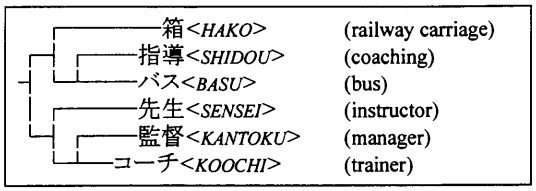
5.4.5 Comparison with alternative methods

We also compared our method with two alternatives, i.e., monolingual distributional clustering (mentioned in Subsection 5.2.2) and naive translingual clustering (mentioned in Subsection 5.2.3), and concluded that it outperforms the alternatives. Example results are shown in Fig. 5.8. The target word was "race," and 17 translation equivalents rendered by the EDR bilingual dictionary and that occur ten or more times in the corpus were clustered. While our method excluded the corpus-irrelevant translation equivalents, the alternatives output dendrograms including all the translation equivalents. Note that deleting the corpus-irrelevant translation equivalents from the dendrograms output by the alternatives would not result in appropriate ones.

5.5 Discussion

5.5.1 Advantages of proposed method

Our method has several practical advantages. First, it produces a corpus-dependent inventory of word senses, eliminating subjective judgment. The resulting inventory covers most senses relevant to a domain, while it excludes senses irrelevant to the domain. In addition, the senses are defined by their translation equivalents appearing in



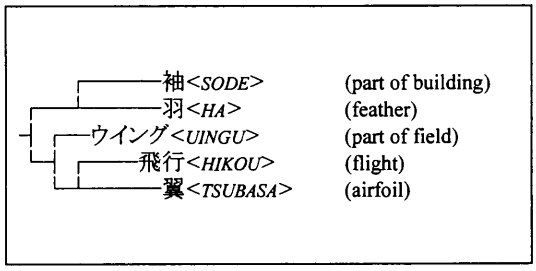
(EDR dictionary)

Translation equivalents that occur ten or more times in the corpus are 指導<*SHIDOU*>, バス<*BASU*>, 監督<*KANTOKU*>, 箱<*HAKO*>, コーチ<*KOOCHI*>, 手引き<*TEBIKI*>, 馬車<*BASHA*>, 客車<*KYAKUSHAI*>.

(Alternative dictionaries)

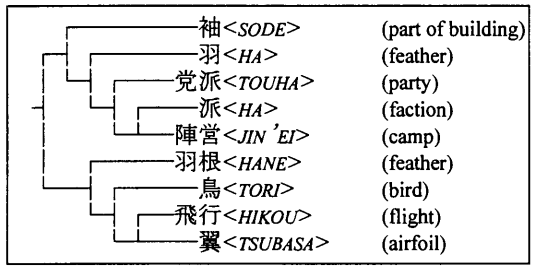
Translation equivalents that occur ten or more times in the corpus are バス<*BASU*>, 監督<*KANTOKU*>, コーチ<*KOOCHI*>, エコノミークラス<*EKONOMII-KURASU*>, 家庭教師<*KATEI-KYOUSHI*>, 客車<*KYAKUSHAI*>.

(a) "coach"



(EDR dictionary)

Translation equivalents that occur ten or more times in the corpus are 腕<*UDE*>, 枝<*EDA*>, 翼<*TSUBASA*>, 飛行<*HIKOU*>, 羽<*HA*>, 飛翔<*HISHOU*>, ウイング<*UINGU*>, 袖<*SODE*>.



(Alternative dictionaries)

Translation equivalents that occur ten or more times in the corpus are(周辺<*SHUUHEN*>, 鳥<*TORI*>, 腕<*UDE*>, 党派<*TOUHA*>, 派<*HA*>, 翼<*TSUBASA*>, 陣営<*JIN 'EI*>, 飛行<*HIKOU*>, 羽<*HA*>, 羽根<*HANE*>, コカイン<*KOKAIN*>, 帆<*HO*>, 袖<*SODE*>.

(b) "wing"

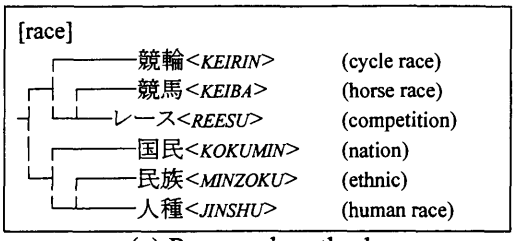
[Note] Alternative dictionaries are *Kenkyusha's English-Japanese Dictionary for the General Reader* and *Kenkyusha's New Collegiate English-Japanese Dictionary*.

Figure 5.7 Clustering results using alternative bilingual dictionaries to select translation equivalents for two target words

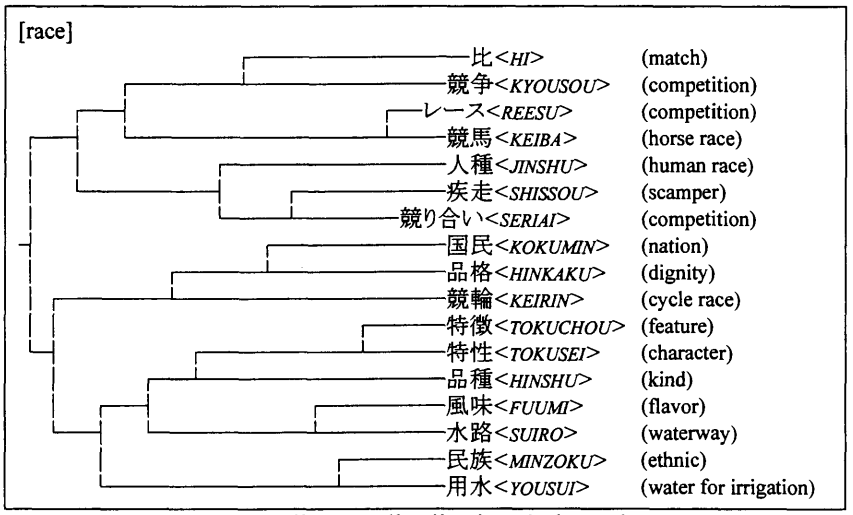
the second-language corpus.

Second, our method unifies word sense acquisition with word sense disambiguation. The sense-vs.-clue correlation matrix is effectively used for word sense disambiguation, as shown in Chapter 4. Therefore, our method guarantees that acquired senses are machine distinguishable, and further it demonstrates the possibility of automatically optimizing the granularity of word senses.

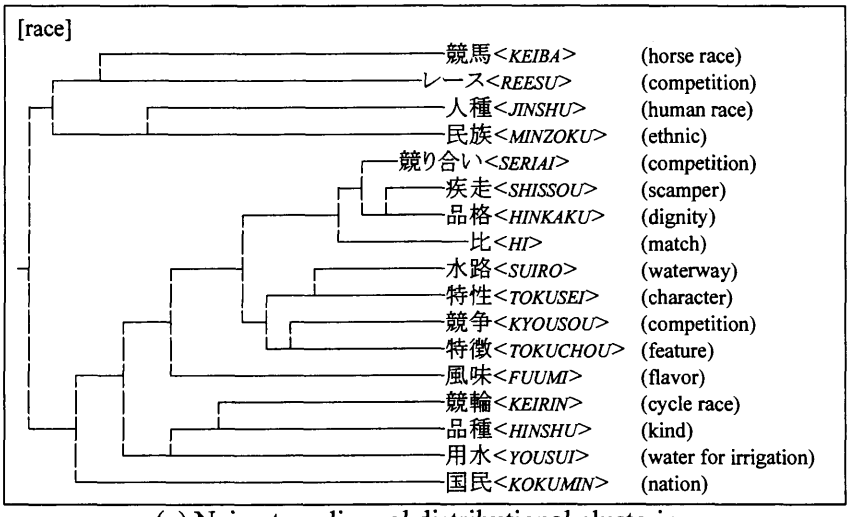
Third, the clustering algorithm is robust to sparse data; it works well not only for high-frequency translation equivalents but also for moderate-frequency translation equivalents due to the iterative calculation of sense-vs.-clue correlations. The moderate-frequency translation equivalents have sparse distribution patterns in the corpus of their own language. However, they are characterized by translingual distribution patterns that are as dense as those characterizing the high-frequency



(a) Proposed method



(b) Monolingual distributional clustering



(c) Naive translingual distributional clustering

Figure 5.8 Clustering with proposed method and two alternative methods

translation equivalents. The experiment showed that translation equivalents with a frequency above ten can be effectively clustered.

Fourth, the computational load is not prohibitive, because a limited number of words, i.e., a few dozen translation equivalents, are clustered for each target word.

While each cycle of clustering includes iterative calculation of sense-vs.-clue correlation matrices, the matrices are relatively small, i.e., a few dozen and less senses by a few hundred clues, and the correlations converge rapidly, as illustrated in Fig. 4.5. Measurement in the experiment described in Section 5.4 showed that it took 35 seconds per target word on a Windows 2000 server (CPU: Pentium 4, clock frequency: 1.9 GHz, memory: 2 GB) to produce a hierarchy of clusters of translation equivalents.

5.5.2 How to evaluate results of word sense acquisition

There is no established or standard method for evaluating the results of word sense acquisition. We proposed two evaluative measures, i.e., the recall of senses and accuracy of sense definitions. However, they have a number of limitations. Evaluating the recall of senses requires a complete list of senses that are used in the corpus. We were thus forced to do evaluation for a limited number of target words. Although it is a good indicator of the overall appropriateness of the clustering results, the accuracy of sense definitions does not specify which senses were defined appropriately.

In their work on word sense acquisition, in which word sense is defined as a set of synonyms, Pantel and Lin (2002) evaluated recall and precision by mapping acquired sets of synonyms to WordNet synsets (Miller 1990). To evaluate the recall, they substituted the complete set of senses to be acquired with pooled results of alternative methods. Therefore, their measure of recall is meaningful only in relation to the alternative methods. In addition, acquired sets of synonyms were mapped to WordNet synsets based on similarity defined ad hoc, and recall and precision depend on the threshold for similarity.

5.5.3 Limitations and directions for extension

There are limitations to the present method. First, while it can produce a hierarchy of clusters, it cannot produce a set of disjoint clusters. Therefore, human decision-making is still required. It is very important to terminate the clustering procedure autonomously during an appropriate cycle, in other words, to determine how many senses are appropriate for the target word. Comparing distribution patterns (not subordinate ones) may be useful for terminating merging; senses characterized by complementary distribution patterns should not be merged.

Second, the present algorithm assumes that each translation equivalent represents one and only one sense of the target word, but this is not always true even when the language pair is English and Japanese. A Japanese *katakana* word resulting from transliteration of an English word sometimes represents multiple senses of the English

word. The algorithm should be extended so that it can detect and split translation equivalents representing more than one sense of the target word.

Third, not only are acquired senses rather coarse-grained, generic senses are difficult to acquire. One of the reasons for this may be that we rely on topically related words. The fact that most distributional word clustering methods use syntactically related words (Hindle 1990; Pereira, et al. 1993; Grefenstette 1994; Lin 1998) suggests that using syntactically related words could improve the performance of our method.

5.6 Related work

Fukumoto and Tsujii (1994) addressed recognition of verbal polysemy. For that purpose, they developed a distributional word-clustering algorithm that is capable of producing overlapping clusters. Verbs are characterized by vectors of the nouns co-occurring with them, and each polysemous verb and its synonyms are clustered. As a result, a polysemous verb is assigned to two or more clusters, each of which corresponds to one of its senses. Fukumoto and Tsujii dealt with polysemy of common verbs, which seems the most difficult one, and their work did not proceed beyond the preliminary evaluation stage.

Pantel and Lin (2002) developed an efficient overlapping clustering algorithm, called CBC (clustering by committee), and applied it to acquiring the word senses of English nouns. CBC initially extracts a set of tight clusters called committees and then assigns words to their most similar clusters. After assigning a word to a cluster, features shared with the cluster are removed from the word. This allows a word to be assigned to multiple clusters, each of which represents one of its senses. CBC can be applied to a set consisting of a large number of words, not only to a set consisting of a polysemous word and its synonyms. They clustered 13,403 nouns extracted from a 1-GB newspaper article corpus, resulting in 941 clusters of synonyms. The recall and precision of senses were 50.8% and 60.8%, respectively.

Schuetze's (1998) method for word sense discrimination, cited in Section 4.7, is also categorized into work on word sense acquisition. He divided the problem of word sense disambiguation into two subproblems: sense discrimination and sense labeling. Sense discrimination divides the occurrences of a word into a number of classes by determining for any two occurrences whether they belong to the same sense or not. His proposed algorithm groups the occurrences of a polysemous word into clusters that consist of contextually similar occurrences. He was not concerned with the subproblem of sense labeling. Therefore, his method acquires sense divisions, but it does not produce definitions of senses such as sets of synonyms and sets of translation

equivalents. He did an experiment in which, for each of 10 test polysemous words, more than a 1000 occurrences (documents) were grouped into either two or ten clusters. The accuracy was 80.8% (two clusters) and 83.1% (ten clusters).

Recently, Widdows and Dorow (2002) proposed another method for discriminating word senses without labeling. It is based on a graph model, and it incrementally builds clusters consisting of semantically related words.

All these methods are based on the clustering technique. From the computational point of view, our method has advantage over the others; it is robust to sparse data and relatively efficient, as discussed in Subsection 5.5.1. The other methods do not have a special measure for the data sparseness problem. Pantel and Lin's method is effective only for words that occur frequently in a large corpus. Moreover, it needs to cluster a huge number of words, each of which is characterized by several hundreds features. Schuetze's method needs to cluster a large number of documents, each of which is represented by a word vector with a few thousand dimensions, for each target word.

It is premature to compare the quality of word senses acquired by different methods. Instead, a few comments on the capabilities of different approaches are given below. First, all of the present methods cannot determine the appropriate number of senses for each target word. Pantel and Lin's method often assigns a word to two or more clusters that actually represent the same sense. Schuetze's experiment was done by restricting the number of clusters to either two or ten, and, in the case of ten clusters, multiple clusters corresponding to the same sense were produced. Our method does not produce a set of disjoint clusters, but a hierarchy of clusters. Second, Pantel and Lin's method acquires senses of all target words through a single run of the clustering, while the other methods, including ours, execute the clustering for each target word. While a single run for all target words is advantageous from the viewpoint of computational efficiency, it prevents the optimum set of senses being acquired for each target word. In contrast, one run per target word can potentially acquire senses with the optimum granularity for each target word.

5.7 Summary

We developed a method for dividing and defining the senses of a polysemous word based on a bilingual comparable corpus. It clusters translation equivalents of the target polysemous word hierarchically based on the similarity of translingual distribution patterns obtained by using the sense-vs.-clue correlation calculation method described in Chapter 4. In each cycle of the clustering procedure, sense-vs.-clue correlations are calculated not only for the full set of senses but also for the sets of senses excluding one

of these senses, so that distribution patterns for senses defined by less dominant translation equivalents can be elicited effectively.

This translingual distributional clustering has several advantages over conventional monolingual distributional clustering. First, it characterizes each translation equivalent by the distribution pattern for its sense relevant to the target word, so it prevents polysemous translation equivalents from degrading the clusters. Second, it works well not only for high-frequency translation equivalents but also mid-frequency ones because the sense-vs.-clue correlation calculation smoothes out the sparse data. Third, the computational load is moderate because it clusters a limited number of translation equivalents, each of which is characterized by a distribution pattern with a relatively small number of dimensions.

The effectiveness of this method was demonstrated through an experiment using Wall Street Journal and Nihon Keizai Shimbun corpora and the EDR bilingual dictionary. The recall of senses was 87.1% for senses whose ratio in the corpus was not less than 5%, and the accuracy of sense definitions was 76.6%.

Chapter 6

Conclusion

6.1 This work

We addressed automatic construction of semantic lexicons from corpora. Specifically, we developed an innovative method for producing a word sense association network from a bilingual comparable corpus and a bilingual dictionary.

In Chapter 2, we described a word sense association network consisting of nodes, each of which represents a sense defined as a set of bilingual synonyms, and edges, each of which connects a pair of associated senses. Then, we presented a framework for producing a word sense association network automatically. The basic idea is to align word association networks in two languages, which are produced from the respective language part of a bilingual comparable corpus, using a bilingual dictionary. The framework consists of translation equivalent extraction, sense-vs.-clue correlation calculation, translation equivalent clustering, and word sense unification.

In Chapter 3, we described a method for extracting pairs of translation equivalents from unaligned bilingual corpora. Presupposing a bilingual dictionary of basic words, it extracts new pairs of translation equivalents by evaluating context similarity between words in the two languages. The method has advantages specific to linguistic and to statistical methods; it does not require a very large corpus, and it can extract not only pairs of simple words but also pairs of compound words and mixed pairs of simple and compound words. The effectiveness of the new method was demonstrated through an experiment using Japanese and English patent specification documents.

In Chapter 4, we described a method for calculating correlations between the senses of a polysemous word and the clues identifying the sense of that word based on the translingual alignment of word associations. The newly developed iterative algorithm, which is based on the assumption that the correlation between a sense and a clue depends on those between the sense and related clues, overcomes the difficulty caused by a disparity in topical coverage between corpora of different languages as well as the data sparseness problem. This method can be regarded as a fully unsupervised learning method for word sense disambiguation. Its effectiveness was demonstrated through a word sense disambiguation (WSD) experiment using Wall Street Journal and Nihon

Keizai Shimbun corpora and the EDR bilingual dictionary.

In Chapter 5, we described a method for clustering translation equivalents of a polysemous word to divide and define its senses corpus-dependently. It calculates sense-vs.-clue correlations using the method mentioned above to characterize (clusters of) translation equivalents by their translingually aligned distribution patterns. This translingual distributional clustering overcomes the data sparseness problem as well as the problem of polysemous translation equivalents and corpus-irrelevant translation equivalents. The feasibility of the method was demonstrated through an experiment using the same corpora and bilingual dictionary used in the WSD experiment.

We have thus shown that a domain-dependent word sense association network can be produced automatically. This network can serve as the basis for semantics-oriented natural language processing. Its applicability to weakly comparable corpora, which are widely available, makes the method attractive.

6.2 Future work

Although we demonstrated its feasibility and effectiveness, our approach needs to be further extended. A major reason for its limited performance is that we do not parse the sentences in the corpora. Our next step will be to apply parsing technology, which has greatly advanced in recent years (Uszkoreit 2002), to word sense acquisition and disambiguation as well as to translation equivalent extraction, which should improve their performance. Syntactic co-occurrence has been effectively used in monolingual distributional word clustering and should be effective for translingual distributional word clustering. In word sense disambiguation, different types of clues are appropriate for different types of polysemous words. Obviously, syntactically related words are useful for disambiguating certain kinds of words. As for translation equivalent extraction, the use of parsing technology will be useful in extracting compound words accurately.

We also plan to demonstrate the benefit of the word sense association network for a variety of natural language processing tasks. For example, we are developing a front-end processor for Web search engines that will use the word sense association network to help the user enter an unambiguous query. It will then expand the query using bilingual synonyms, and the expanded query will then be sent to the search engine, which can be one for a language other than that used to enter the query. The retrieved documents will be filtered by disambiguating the query terms they contain. The word sense association network will thus play essential roles at both ends of the process.

Acknowledgments

I would like to express my sincere appreciation to Professor Toyoaki Nishida for his kind guidance, intensive reviewing, and valuable comments, all of which enabled me to complete this thesis. I would also like to thank Professor Jun Adachi, Professor Keikichi Hirose, Professor Mitsuru Ishizuka, Professor Sadao Kurohashi, Professor Hiroshi Nakagawa, and Professor Junichi Tsujii for their kind reviews and insightful comments, which provided me with invaluable feedback. It is worth mentioning specially that Professor Tsujii's enlightening discussions and comments have had an impact on my research that goes beyond the work described in this thesis.

I would like to express my special gratitude to Professor Makoto Nagao for the valuable advice and continuous encouragement that have kept me focused on research of natural language processing over the last two decades.

The work described in this thesis was conducted at Hitachi's Central Research Laboratory starting in the mid-1990s. I would like to thank Dr. Michiharu Nakamura, Dr. Eiji Takeda, and Dr. Toshikazu Nishino for providing me with the opportunity to conduct the research and for providing an excellent research environment. I am also grateful to my colleagues who kindly extended advice, discussion, and assistance: Toshiko Aizono, Naoto Akira, Dr. Hiromichi Fujisawa, Dr. Toru Hisamitsu, Dr. Osamu Imaichi, Dr. Taizo Kinoshita, Atsuko Koizumi, Hiroyuki Maezawa, Junichi Matsuda, Yasutsugu Morimoto, Akiyoshi Nakahara, Yoshito Nejime, Dr. Yoshiki Niwa, Dr. Mamoru Sugie, Dr. Norihiro Suzuki, Prof. Ryuichi Suzuki, and Dr. Tetsuo Yokoyama.

The research on which this thesis is based was in part supported by the Telecommunications Advancement Organization of Japan and the New Energy and Industrial Technology Development Organization of Japan.

The work I have described originated from my experience in researching and developing a machine translation system in the Systems Development Laboratory and at Software Works, Hitachi, Ltd. in the 1980s. I am grateful to the people who gave me their kind advice and support during that period: Prof. Koichiro Ishihara, Hiroshi Isobe, Teiichi Kashiwagi, Dr. Jun Kawasaki, Prof. Zenshiro Kawasaki, Atsu Kimura, Keiichi Matsumoto, Prof. Yoshihiko Nitta, Dr. Sakae Takahashi, Prof. Tan Watanabe, Noriyuki Yamasaki, and Kikuo Yoshimura.

In addition, I would like to express my appreciation to Professor Toshiyuki Sakai and Professor Takeo Kanade for the significant role they played in my development as a researcher as well as to Dr. Takeo Miura and Dr. Koichi Haruna for giving me the

chance to enter the interesting and challenging field of natural language processing.

Finally, I am sincerely grateful for the constant support and encouragement given by my wife Kazuyo.

References

Agirre, Eneko and German Rigau. 1996. Word sense disambiguation using conceptual density. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 16-22.

Agirre, Eneko, Olatz Ansa, David Martinez, and Eduard Hovy. 2001. Enriching WordNet concepts with topic signatures. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*.

Allegrini, Paolo, Simonetta Montemagni, and Vito Pirrelli. 2000. Learning word clusters from data types. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 8-14.

Basili, Roberto, Michelangelo Della Rocca, and Maria Tereza Pazienza. 1997. Towards a bootstrapping framework for corpus semantic tagging. In *Proceedings of the ACL-SIGLEX Workshop "Tagging Text with Lexical Semantics: Why, What, and How?"* pages 66-73.

Black, Ezra. 1988. An experiment in computational discrimination of English word senses. *IBM Journal of Research and Development*, 32(2): 185-194.

Brill, Eric, David Magerman, Mitchell Marcus, and Beatrice Santorini. 1990. Deducing linguistic structure from the statistics of large corpora. In *Proceedings of DARPA Speech and Natural Language Workshop*.

Brown, Peter F., Jennifer C. Lai, and Robert L. Mercer. 1991a. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 169-176.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991b. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 264-270.

Brown, Peter F., Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4): 467-479.

Bruce, Rebecca and Janyce Wiebe. 1994. Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 139-145.

Calzolari, Nicoletta and Remo Bindi. 1990. Acquisition of lexical information from a large textual Italian corpus. In *Proceedings of the 13th International Conference on Computational Linguistics*, Vol. 3, pages 54-59.

Chen, Stanley F. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 9-16.

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1): 22-29.

Church, Kenneth W. and Robert L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1): 1-24.

Clark, Stephen and David Weir. 2000. A class-based probabilistic approach to structural disambiguation. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 194-200.

Cowie, Jim, Joe Guthrie, and Louisa Guthrie. 1992. Lexical disambiguation using simulated annealing. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 359-365.

Dagan, Ido and Alon Itai. 1990. Automatic acquisition of constraints for the resolution of anaphora references and syntactic ambiguities. In *Proceedings of the 13th International Conference on Computational Linguistics*, Vol. 3, pages 330-332.

Dagan, Ido, Kenneth W. Church, and William A. Gale. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of Workshop on Very Large Corpora*, pages 1-8.

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

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, Series B, 39: 1-38.

Dolan, William B. 1994. Word sense ambiguation: clustering related senses. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 712-716.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1): 61-74.

EDR. 1990a. *Concept Dictionary*, Technical Report TR-027, Japan Electronic Dictionary Research Institute, Ltd, Tokyo.

EDR. 1990b. *Bilingual Dictionary*, Technical Report TR-029, Japan Electronic Dictionary Research Institute, Ltd, Tokyo.

Fellbaum, Christiane (ed.) 1998. *WordNet: An electronic lexical database,* The MIT Press, Cambridge, MA.

Fukumoto, Fumiyo and Junichi Tsujii. 1994. Automatic recognition of verbal polysemy. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 762-768.

Fung, Pascale. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 236-243.

Fung, Pascale and Kathleen McKeown. 1997. Finding terminology translations from

non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192-202.

Fung, Pascale and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 414-420.

Gale, William A. and Kenneth W. Church. 1991a. Identifying word correspondences in parallel texts. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pages 152-157.

Gale, William A. and Kenneth W. Church. 1991b. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 177-184.

Gale, William A., Kenneth W. Church, and David Yarowsky. 1992a. Estimating upper and lower bounds on the performance of word-sense disambiguation programs, In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 249-256.

Gale, William A., Kenneth W. Church, and David Yarowsky. 1992b. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 101-112.

Gale, William A., Kenneth W. Church, and David Yarowsky. 1993. A method for disambiguating word senses in large corpus. *Computers and the Humanities*, 26: 415-439.

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Boston.

Guthrie, Joe A., Louise Guthrie, Yorick Wilks, and Homa Aidinejad. 1991. Subject-dependent co-occurrence and word sense disambiguation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 146-152.

Harabagiu, Sanda M., George A. Miller, and Dan I. Moldovan. 1999. WordNet 2 - A morphologically and semantically enhanced resource. In *Proceedings of the ACL-SIGLEX Workshop "Standardizing Lexical Resources,"* pages 1-8.

Harris, Zellig. 1985. Distributional structure. In: Jerrold Katz (ed.) *The Philosophy of Linguistics*, Oxford University Press, New York, pages 26-47.

Hearst, Marti A. 1991. 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*, pages 1-22.

Hearst, Marti A. 1992. Automatic acquisition of hyponym from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 539-545.

Hindle, Donald. 1990. Noun classification from predicate-argument structures. In

Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics, pages 268-275.

Hindle, Donald and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1): 103-120.

Ide, Nancy and Jean Veronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1): 1-40.

Inoue, Naomi and Izuru Nogaito. 1993. Automatic construction of the Japanese-English dictionary from bilingual text, *Technical Report of IEICE*, NLC93-39 (in Japanese).

Ioannis-Dimitrios, Koutsoubos and Christodoulakis Dimitris. 2002. Requirements for domain-specific WordNets. In *Proceedings of Workshop on WordNet Structures and Standardization, and how these affect WordNet Applications and Evaluation, the 3rd International Conference on Language Resources and Evaluation*, pages 52-55.

Ishimoto, Hiroyuki and Makoto Nagao. 1994. Automatic construction of a bilingual dictionary of technical terms from parallel texts, *Technical Report of IPSJ*, NL-102-11 (in Japanese).

Jain, A. K., M. N. Murty, and P. J. Flynn. 1999. Data clustering: a review. *Computing Surveys*, 31(3): 264-323.

Kaji, Hiroyuki and Toshiko Aizono. 1996. Extracting word correspondences from bilingual corpora based on word co-occurrence information. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 23-28.

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.

Karov, Yael and Shimon Edelman. 1998. Similarity-based word sense disambiguation. *Computational Linguistics*, 24(1): 41-59.

Kay, Martin and Martin Roscheisen. 1993. Text-translation alignment, *Computational Linguistics*, 19(1): 121-142.

Kelly, Edward F. and Philip J. Stone. 1975. *Computer Recognition of English Word Senses*, North-Holland, Amsterdam.

KENKYUSHA. 1984. *Kenkyusha's English-Japanese Dictionary for the General Reader (First edition)*, Edited by Tokuichiro Matsuda, Ichiro Yokoyama, and Nobuyuki Higashi, Kenkyusha Ltd., Tokyo.

KENKYUSHA. 1985. *Kenkyusha's New Collegiate English-Japanese Dictionary (Fifth edition)*, Edited by Yoshio Koine, Kikuo Yamakawa, Shigeru Takebayashi, and Michio Yoshikawa, Kenkyusha Ltd., Tokyo.

Kikui, Genichiro. 1998. Term-list translation using mono-lingual word co-occurrence vectors. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 670-674.

Kilgarriff, Adam. 1998. "I don't believe in word senses." *Computers and the Humanities*, 31(2): 91-113.

Kilgarriff, Adam and Joseph Rosenzweig. 2000. English SENSEVAL: report and results. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Vol. 3, pages 1239-1244.

Kilgarriff, Adam. 2001. English lexical sample task description. In *Proceedings of Senseval 2, the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 17-20.

Kitamura, Mihoko and Yuji Matsumoto. 1996. Automatic extraction of word sequence correspondences in parallel corpora, In *Proceedings of the 4th Workshop on Very Large Corpora*, pages 79-87.

Kumano, Akira and Hideki Hirakawa. 1994. Building an MT dictionary from parallel texts based on linguistic and statistical information, In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 76-81.

Kupiec, Julian. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 17-22.

LDOCE. 1995. *Longman Dictionary of Contemporary English (Third edition)*, Edited by Della Summers, Pearson Education Ltd., Harlow.

Leacock, Claudia, Geoffrey Towell, and Ellen Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings of the ARPA Human Language Technology Workshop*, Morgan Kaufman.

Lesk, Michael E. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th International Conference on Systems Documentation*, pages 24-26.

Li, Hang and Naoki Abe. 1998. Word clustering and disambiguation based on co-occurrence data. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 749-755.

Li, Xiaobin, Stan Szpakowicz, and Stan Matwin. 1995. A WordNet-based algorithm for word sense disambiguation. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1368-1374.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 768-774.

Lin, Dekang and Patrick Pantel. 2002. Concept discovery from text. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 577-583.

Luk, Alpha K. 1995. Statistical sense disambiguation with relatively small corpora using dictionary definitions. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 181-188.

Melamed, I. Dan. 1997a. A portable algorithm for mapping bitext correspondence. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for*

Computational Linguistics, pages 305-312.

Melamed, I. Dan. 1997b. A word-for-word model of translational equivalence. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 490-497.

Mihalcea, Rad and Dan Moldovan. 1999. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 152-158.

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

Miller, George A., Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 240-243.

Nakagawa, Hiroshi. 2001. Disambiguating of compound noun translations extracted from bilingual comparable corpora. In *Proceedings of the 6th Natural Language Processing Pacific-Rim Symposium*, pages 67-74.

Niwa, Yoshiki and Yoshihiko Nitta. 1994. Co-occurrence vectors from corpora vs. distance vector from dictionaries. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 304-309.

Pantel, Patrick and Dekang Lin. 2002. Discovering word senses from text, In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613-619.

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.

Rapp, Reinhard. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 320-322.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 320-322.

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.

ROGET. 1977. *Roget's International Thesaurus (Fourth edition)*, Revised by Robert L. Chapman, Thomas Y. Crowell, Publishers, New York and Harper & Row, Publishers,

London.

Ruge, Gerda. 1991. Experiments on linguistically based term associations. In *Proceedings of RIAO '91, Conference on Intelligent Text and Image Handling,* pages 528-545.

Schuetze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1): 97-124.

SENSEVAL-2. 2001. *Proceedings of Senseval 2, the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems*, Edited by Judita Preiss and David Yarowsky, The Association for Computational Linguistics, New Brunswick, NJ.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1): 143-177.

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.

Ushioda, Akira. 1996. Hierarchical clustering of words and applications to NLP tasks. In *Proceedings of the 4th Workshop on Very Large Corpora*, pages 28-41.

Uszkoreit, Hans. 2002. New chances for deep linguistic processing. In *Proceedings of the 19th International Conference on Computational Linguistics.*

van Rijsbergen, C. J. 1979. *Information retrieval (2nd edition)*, Butterworths, London.

Veronis, Jean and Nancy Ide. 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 389-394.

Vossen, Piek (ed.) 1998. *EuroWordNet: A multilingual database with lexical semantic networks,* Kluwer Academic Publishers, Dordrecht.

Walker, Donald E. and Robert A. Amsler. 1986. The use of machine-readable dictionaries in sublanguage analysis. In *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, Edited by Ralph Grishman and Richard Kittredge, L. Erlbaum Associates, Hillsdale, NJ, pages 69-84.

Weiss, S. 1973. Learning to disambiguate. *Information Storage and Retrieval*, 9: 33-41.

Widdows, Dominic and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1093-1099.

Yamamoto, Yukio and Masasi Sakamoto. 1993. Extraction of technical term bilingual dictionary from bilingual corpus, *Technical Report of IPSJ*, NL-94-12 (in Japanese).

Yarowsky, David. 1992. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 454-460.

Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189-196.

Yokoi, Toshio. 1995. The EDR electronic dictionary. *Communication of the ACM*, 38(11): 42-44.

Zernik, Uri. 1991. Train1 vs. Train2: tagging word senses in a corpus. In *Proceedings of RIAO'91, a Conference on Intelligent Text and Image Handling*, pages 567-585.

Publication List

Refereed Journal Papers

Kaji, Hiroyuki and Yoshihiko Nitta. 1979. A fuzzy model of a document retrieval system and its implementation. *Trans. IECE*, 62-D(4): 297-304 (in Japanese) and *Systems, Computers and Controls* (Scripta Publishing Co.), 10(2): 76-85 (English translation version).

梶博行, 新田義彦, "文献検索システムのファジイモデルとその実現," 電子通信学会論文誌, 62-D(4): 297-304 (1979 年 4 月).

Haruna, Koichi, Kazuo Nakao, Norihisa Komoda, and Hiroyuki Kaji. 1983. A method to develop objectives tree —PPDS—. *Keisoku to Seigyo*, 22(2): 185-200 (in Japanese).

春名公一, 中尾和夫, 薦田憲久, 梶博行, "目的樹木作成技法－PPDS－の開発," 計測と制御, 22(2): 185-200 (1983 年 2 月).

Kaji, Hiroyuki, Atsuko Koizumi and Kikuo Yoshimura. 1989. A semantics-based machine translation system from Japanese into English. *Future Computing Systems* (Oxford University Press), 2(3): 247-259.

Hirai, Akihiro, Hiroyuki Kaji, and Minoru Ashizawa. 1990. Ambiguity detection in Japanese dependency structures for pre-editing. *Tran. IPSJ*, 31(10): 1425-1437 (in Japanese).

平井章博, 梶博行, 芦沢実, "機械翻訳向け前編集のための日本語係り受け構造の曖昧性検出方式," 情報処理学会論文誌, 31(10): 1425-1437 (1990 年 10 月).

Kaji, Hiroyuki, and Toshiko Aizono. 2001. Extracting word translations from bilingual corpora based on similarity of co-occurring word sets. *Tran. IPSJ*, 42(9): 2248-2258 (in Japanese).

梶博行, 相薗敏子, "共起語集合の類似度に基づく対訳コーパスからの対訳語抽出," 情報処理学会論文誌, 42(9): 2248-2258 (2001 年 9 月).

Kaji, Hiroyuki, Yasutsugu Morimoto, and Toshiko Aizono. 2003. Extracting a topic hierarchy from a text corpus. *Tran. IPSJ*, 44(2): 405-420 (in Japanese).

梶博行, 森本康嗣, 相薗敏子, "テキストコーパスからのトピック階層の抽出," 情報処理学会論文誌, 44(2): 405-420 (2003 年 2 月).

Refereed International Conference Papers

Nakao, Kazuo, Koichi Haruna, Norihisa Komoda and Hiroyuki Kaji. 1980. A structural approach to system requirements analysis of information systems. In *Proceedings of the 4th International Computer Software & Applications Conference*, pages 207-213.

Komoda, Norihisa, Koichi Haruna, Hiroyuki Kaji and Hiroshi Shinozawa. 1981. An innovative approach to system requirements analysis by using structural modeling method. In *Proceedings of the 5th International Conference on Software Engineering*, pages 305-313.

Kaji, Hiroyuki, Norihisa Komoda, Koichi Haruna, Hiroyuki Kitajima, and Kazuo Nakao. 1981. An interactive system for analyzing complex system structure. In *Proceedings of International Conference on Cybernetics and Society*, pages 213-217.

Nitta, Yoshihiko, Atsushi Okajima, Hiroyuki Kaji, Youichi Hidano, and Koichiro Ishihara. 1984. A proper treatment of syntax and semantics in machine translation. In *Proceedings of the 10th International Conference on Computational Linguistics*, pages 159-166.

Kaji, Hiroyuki. 1988. An efficient execution method for rule-based machine translation. In *Proceedings of the 12th International Conference on Computational Linguistics*, pages 824-829.

Kaji, Hiroyuki, Yuko Kida, and Yasutsugu Morimoto. 1992. Learning translation templates from bilingual text. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 672-678.

Kaji, Hiroyuki and Toshiko Aizono. 1996. Extracting word correspondences from bilingual corpora based on word co-occurrence information. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 23-28.

Kaji, Hiroyuki, Yasutsugu Morimoto, Toshiko Aizono, and Noriyuki Yamasaki. 1999. Navigation in an association thesaurus automatically generated from a corpus. In *Proceedings of Workshop on Text Mining, the 16th International Joint Conference on Artificial Intelligence*, pages 64-74.

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

Kaji, Hiroyuki and Yasutsugu Morimoto. 2002. Towards sense-disambiguated association thesauri. In *Proceedings of LREC 2002 (the 3rd International Conference on Language Resources and Evaluation) Workshop "Using Semantics for Information Retrieval and Filtering,"* pages 59-64.

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.

Journal Articles

Kaji, Hiroyuki, and Atsushi Okajima. 1985. Machine translation systems in Hitachi. *Joho Shori*, 26(10): 1214-1216 (in Japanese).

梶博行, 岡島惇, "日立における機械翻訳システム," 情報処理, 26(10): 1214-1216 (1985年10月).

Kaji, Hiroyuki and Hiromichi Fujisawa. 1996. Technologies for digital library systems. *Journal of IEICE*, 79(9): 910-919 (in Japanese).

梶博行, 藤澤浩道, "電子図書館システムの技術動向," 電子情報通信学会誌, 79(9): 910-919 (1996年9月).

Other International Conference Papers

Nitta, Yoshihiko and Hiroyuki Kaji. 1977. A practical approach to pictorial data retrieval. *Workshop on Pattern Database Systems* (Tokyo).

Kaji, Hiroyuki. 1987. HICATS/JE: A Japanese-to-English machine translation system based on semantics. In *Proceedings of Machine Translation Summit*, pages 55-60.

Kaji, Hiroyuki. 1988. Current status of machine translation. *Language and Computer Symposium* (Jakarta).

Kaji, Hiroyuki. 1988. Dictionary structure for flexible lexical transfer in machine translation. In *Proceedings of International Symposium on Electronic Dictionaries*, pages 36-38.

Kaji, Hiroyuki. 1989. Language control for effective utilization of HICATS/JE. In *Proceedings of Machine Translation Summit II*, pages 72-77.

Kaji, Hiroyuki. 1989. Problems in natural language generation for machine translation. *Workshop on MT Basic Research into the '90s* (Manchester).

Kaji, Hiroyuki, Yuko Kida and Yasutsugu Morimoto. 1993. Learning translation templates from bilingual text. *French-Japanese Workshop on Machine Translation* (Tokyo).

Morimoto, Yasutsugu, Toshiko Aizono, and Hiroyuki Kaji. 1998. Generation of a corpus-dependent thesaurus and interactive text retrieval. In *Proceedings of JSPS-HITACHI Workshop on New Challenges in Natural Language Processing and its Application*, pages 65-68.

Kaji, Hiroyuki, Yasutsugu Morimoto, and Toshiko Aizono. 1998. Acquiring translation knowledge from bilingual corpora. In *Proceedings of JSPS-HITACHI Workshop on New Challenges in Natural Language Processing and its Application*, pages 88-93.

Kaji, Hiroyuki, Yasutsugu Morimoto, and Toshiko Aizono. 1998. Automatic thesaurus generation from corpora for text retrieval. *German-Japanese Workshop on Lexical Resources for Information Retrieval* (Stuttgart).

Morimoto, Yasutsugu, Toshiko Aizono, Hiroyuki Kaji, and Noriyuki Yamasaki. 1999. A corpus-dependent thesaurus generator/navigator. *The 3rd European Conference on Research and Advanced Technology in Digital Libraries*, Demo Session.

Kaji, Hiroyuki. 1999. Controlled languages for machine translation: State of the art. In *Proceedings of Machine Translation Summit VII*, pages 37-39.

Kaji, Hiroyuki, Yasutsugu Morimoto, Toshiko Aizono, and Noriyuki Yamasaki. 2000. Corpus-dependent association thesauri for text mining. *German-Japanese Workshop on Natural Language Processing and Information Retrieval* (Yokohama).

Kaji, Hiroyuki. 2003. Word sense acquisition from bilingual comparable corpora. *German-Japanese Workshop on NLP for Semantic Web and Information Management* (Sapporo).

Other Conference Papers and Technical Reports (in Japanese)

坂井利之, 金出武雄, 梶博行, "AND-OR 特徴トリーによる認識対象の表現とそのマンマシン対話的生成," 電子通信学会パターン認識と学習研究会技術報告, PRL75-27 (1975 年 7 月).

梶博行, 新田義彦, "Fuzzy 集合論的文献検索システム," 昭和 53 年度電子通信学会総合全国大会講演論文集, 分冊 5, p. 197 (1978 年 3 月).

梶博行, 新田義彦, "マイクロ化技術情報の会話型検索システムにおけるソフトウエアの提案," 昭和 53 年電気学会全国大会講演論文集, pp. 1426-1427 (1978 年 4 月).

梶博行, 薦田憲久, 中尾和夫, 春名公一, "グラフィックディスプレイを用いたシステム構造分析システムの提案," 情報処理学会第 20 回全国大会講演論文集, pp. 669-670 (1979 年 7 月).

梶博行, 中尾和夫, 春名公一, "階層グラフの分割表示のためのアルゴリズムの提案," 情報処理学会第 21 回全国大会講演論文集, pp. 35-36 (1980 年 5 月).

梶博行, 北島弘行, 薦田憲久, 春名公一, "グラフ処理用データベースシステムの試作," 情報処理学会第 22 回全国大会講演論文集, pp. 667-668 (1981 年 3 月).

北島弘行, 梶博行, 薦田憲久, 春名公一, "システム構造化・汎用支援ツール(ストラクチャベースシステム)の提案," 昭和 56 年電気学会全国大会講演論文集 (1981 年 3 月).

新田義彦, 梶博行, "日英機械翻訳システムNEAT82－(1)概念依存図式による日本語文解析," 情報処理学会第 27 回全国大会講演論文集, pp. 1083-1084 (1983 年 10 月).

梶博行, 新田義彦, "日英機械翻訳システムNEAT82－(2)概念依存図式の変換による英文生成," 情報処理学会第 27 回全国大会講演論文集, pp. 1085-1086 (1983 年 10 月).

梶博行, 新田義彦, "概念依存図式からの英文生成," 情報処理学会第 28 回全国大会講演論文集, pp. 907-908 (1984 年 3 月).

新田義彦, 梶博行, “概念依存図式による日本語文の意味のモデル化,” 情報処理学会第 28 回全国大会講演論文集, pp. 909-910 (1984 年 3 月).

梶博行, 伊佐津敦子, “日英機械翻訳のための日本語文の依存構造解析,” 情報処理学会第30回全国大会講演論文集, pp. 1579-1580 (1985 年 3 月).

梶博行, 吉村紀久雄, 臼井孝雄, “日英機械翻訳システムATHENE／Nにおける文法記述言語,” 情報処理学会第 31 回全国大会講演論文集, pp. 1347-1348 (1985 年 9 月).

村上孝也, 唐木武志, 吉村紀久雄, 梶博行, “HICATS／JEの辞書作成支援環境,” 情報処理学会第 33 回全国大会講演論文集, pp. 1757-1758 (1986 年 10 月).

臼井孝雄, 坂本浩一, 沢田覚, 吉村紀久雄, 梶博行, “HICATS／JEのポストエディット支援環境,” 情報処理学会第 33 回全国大会講演論文集, pp. 1759-1760 (1986 年 10 月).

坂本浩一, 臼井孝雄, 吉村紀久雄, 梶博行, “HICATS／JEの文法作成支援環境,” 情報処理学会第 33 回全国大会講演論文集, pp. 1761-1762 (1986 年 10 月).

小泉敦子, 梶博行, 鶴秀夫, “日英機械翻訳のための日本文の格構造モデル,” 情報処理学会第 34 回全国大会講演論文集, pp. 1273-1274 (1987 年 3 月).

松田純一, 梶博行, 臼井孝雄, “機械翻訳における文法評価,” 情報処理学会第 34 回全国大会講演論文集, pp. 1275-1276 (1987 年 3 月).

平井章博, 梶博行, “日英機械翻訳用前編集支援システムに関する一考察,” 情報処理学会第 35 回全国大会講演論文集, pp. 1243-1244 (1987 年 9 月).

小泉敦子, 梶博行, “機械翻訳の中間言語における述語概念素の設定法,” 情報処理学会第 36 回全国大会講演論文集, pp. 1227-1228 (1988 年 3 月).

松田純一, 梶博行, “英文生成における修飾語句の語順決定方式,” 情報処理学会第 36 回全国大会講演論文集, pp. 1233-1234 (1988 年 3 月).

平井章博, 高岡紀子, 梶博行, “日英機械翻訳用前編集支援システム(1)－構文的曖昧性の検出方式－,” 情報処理学会第 36 回全国大会講演論文集, pp. 1229-1230 (1988 年 3 月).

芦沢実, 平井章博, 梶博行, “日英機械翻訳用前編集支援システム(2)－形態素の曖昧性の検出方式－,” 情報処理学会第 36 回全国大会講演論文集, pp. 1231-1232 (1988 年 3 月).

中島弘之, 梶博行, “テキストからの共起関係自動抽出の試み,” 情報処理学会第 38 回全国大会講演論文集, pp. 325-326 (1989 年 3 月).

中島弘之, 梶博行, “対訳テキストを利用した訳語選択のための共起関係の自動抽出,” 情報処理学会第 39 回全国大会講演論文集, pp. 706-707 (1989 年 10 月).

芦沢実, 梶博行, “対話式日英機械翻訳における意味的なあいまい性の提示方法,” 情報処理学会第 40 回全国大会講演論文集, pp. 407-408 (1990 年 3 月).

芦沢実, 梶博行, “問答式日英機械翻訳における例文による深層格のあいまい性の提示方法,” 情報処理学会第 42 回全国大会講演論文集, 第 3 分冊, pp. 21-22 (1991

年 3 月).

小泉敦子, 梶博行, "日英機械翻訳における格フレーム辞書カスタマイズ方式," 情報処理学会第 43 回全国大会講演論文集, 第 3 分冊, pp. 207-208 (1991 年 10 月).

森本康嗣, 梶博行, "語の対訳知識を用いた対訳テキストの文対応付けアルゴリズム," 言語処理学会第1回年次大会論文集, pp. 93-96 (1995 年 3 月).

梶博行, 森本康嗣, 相薗敏子, 山崎紀之, "知的概念検索の研究開発," 次世代電子図書館システム研究開発事業論文集, pp. 49-52 (1998 年 3 月).

梶博行, 森本康嗣, 相薗敏子, 山崎紀之, 飯田恵子, 内田安彦, "コーパス対応の関連シソーラスナビゲーション," 情報処理学会データベースシステム研究会／情報学基礎研究会研究報告, DBS-118-13 / FI-54-13 (1999 年 5 月).

梶博行, 森本康嗣, 相薗敏子, 山崎紀之, 飯田恵子, 内田安彦, "知的概念検索の研究開発," 次世代電子図書館システム研究開発事業論文集, pp. 157-160 (2000 年 3 月).