Extracting a Disambiguated Thesaurus from Parallel Dictionary Definitions

Naohiko URAMOTO
IBM Research, Tokyo Research Laboratory
1623-14 Shimotsuruma, Yamato-shi, Kanagawa-ken 242, Japan
uramoto@trl.vnet.ibm.com

Abstract

This paper describes a method for extracting disambiguated (bilingual) is-a relationships from parallel (English and Japanese) dictionary definitions by using word-level alignment. Definitions have a specific pattern, namely, a "genus term and differentia" structure; therefore, bilingual genus terms can be extracted by using bilingual pattern matching. For the alignment of words in the genus terms, a dynamic programming framework for sentence-level alignment proposed by Gale et al. [6] is used.

1 Introduction

Deeper and less ambiguous knowledge can be obtained by using parallel corpora than by using monolingual corpora. Research on this topic includes studies by Dagan et al. [4], who used parallel corpora for word selection in the target language in machine translation, and Utsuro et al. [17], who applied sample sentences in a English-to-Japanese dictionary to learning of case-patterns. Development of algorithms for sentence alignment in corpora is also a hot issue [2, 6, 10].

In this paper, bilingual sentences are taken from the IBM Dictionary of Computing [9] (originally written in English) and its Japanese translation [3]. Definitions in dictionaries have a restricted structure, namely, *genus term* and *differentia*. Using bilingual pattern matching, we obtain a bilingual pair consisting of an entry word and its genus term. It is assumed that the definitions in the dictionary and its translation have been aligned, since the matching between them is almost one-to-one, and the definitions are separated by entries, which makes it easy to align definitions. However, words in the genus terms for an entry must be aligned.

Most alignment algorithms require an *anchor point* that combines a part of one sentence and a part of sentence of other language. We use the bilingual pattern of the definitions. Use of the pattern makes it possible to align the words in part of a sentence without consulting a dictionary.

There is no doubt that a thesaurus is one of most useful sources of knowledge for semantic processing. The aim of our work is to develop a domain-dependent thesaurus for example-based disambiguation [16]. Much work has been done (for example, by Amsler [1], Klavans et al. [11], Nakamura et al. [14] and Guthrie et al. [7]) on the extraction of thesauruses from monolingual dictionaries such as the Longman Dictionary of Contemporary English (LDOCE) [15]. However, words in the definitions are ambiguous, and consequently it is difficult to get a disambiguated thesaurus. The advantage of using parallel texts is that it reduces the number of ambiguities inherent in each monolingual text, when the sentences in the texts are aligned. In our approach, the use of a bilingual dictionary makes it possible to acquire a set of

pairs of bilingual is-a relationships. The relationships are represented by [English word : Japanese Word] → [English hypernym : Japanese hypernym].

By simple matching using language-dependent patterns that appear in English and Japanese definitions, a genus term for an entry word can be extracted. The genus term consists of multiple words, and the alignment is not always one-to-one. However, in many cases, the order of words in English and Japanese genus terms is the same. Therefore, algorithms for sentence alignment can be applied to the problem. In this paper, the dynamic programming framework developed by Gale et al. [6] is used to align words in the genus term. In our alignment program, two preferences are used to measure the distance of an alignment. One is matching of syntactic categories, and the second is co-occurrence in other parts of the text.

2 Structure of the Definitions in a Parallel Dictionary

As a bilingual corpus, the IBM Dictionary of Computing [9] (written in English) and its Japanese translation [3] are used. Each contains about 10,000 entries for technical terms in the computer domain. Basically, one English definition is translated as one Japanese definition, and it is therefore easy to align sentences.

For example, the definitions of the entry word "active line" in the English version of the dictionary and its translation are as follows:

active line: 通信回線
(EDEF) a telecommunication line that is currently available for transmission of data.
(JDEF) 現在、データ転送に利用できる通信回線

The structure of definitions of on-line dictionaries has been analyzed (for example, in [14, 7, 18]). The main parts of the definition of a word are a *genus term* and a *differentia*. The genus term represents a hypernym of the entry word, and the differentia is used to distinguish the entry from other entries that have the same genus term.

In the English definition (EDEF), "telecommunication line" is the genus term for the entry "active line," and "that is currently available for transmission" is the differentia part. In the Japanese definition (JDEF), "通信回線" is the genus term, while "現在、データ転送に利用できる" is the differentia. The position of the genus term depends on the language. In English, it appears the beginning of the definition, while in Japanese it come at the end of the definition.

The following are the bilingual patterns for extracting the genus terms from the English and Japanese definitions:

(EPAT) PRE-DIFF GENUS-TERM+ POST-DIFF**
(JPAT) PRE-DIFF GENUS-TERM+*

The expression "WORD*" matches zero or more words, and "WORD+" matches one or more words. The expression GENUS-TERM matches words that have same syntactic category as an entry. PRE-DIFF matches a determiner. POST-DIFF matches a sequence that begins with a word whose category is not the same as that of the entry word. Information on the parts of speech of the words in the definitions is needed in order to recognize the genus words by using the patterns.

a_DET telecommunication_NOUN line_NOUN that_REL is_VERB available_ADJ for_PREP transmission_NOUN of_PREP data_NOUN

Figure 1: Tagged English definition for the entry "active line"

現在_ADV 、_PUNC データ転送_NOUN に_KJYO 利用_NOUN できる_JYODO 通信_NOUN 回線_NOUN

Figure 2: Tagged Japanese definition for the entry "active line"

2.1 Extraction of a Disambiguated Thesaurus

The genus term for an entry is extracted by using the bilingual pattern. The extraction procedure has three steps:

1. Part-of-speech tagging of parallel definitions
2. Matching of genus terms by using bilingual pattern
3. Alignment of the words in the genus terms extracted in step 2

First, the parts of speech of the definition sentences are tagged automatically. For English analysis, the English Slot Grammar (ESG) developed by McCord [13] is used. The Japanese Morphological Analyzer (JMA) developed by Maruyama et al. [12] is used for Japanese definitions. Figure 1 and 2 show the outputs of the English and Japanese taggers.

For each tagged parallel definition, the pattern described in Section 1 is applied. The result of matching is as follows:

Matching result of English definition:

ENG-ENTRY = active line
PRE-DIFF-1 = a_DET
GENUS-TERM-1 = telecommunication_NOUN
GENUS-TERM-2 = line_NOUN
POST-DIFF-1 = that_REL
POST-DIFF-2 = is_VERB
(other differentiae)

Matching result of Japanese definition:

JPN-ENTRY = 活動回線
PRE-DIFF-1 = 現在_ADV
PRE-DIFF-2 = データ_NOUN
(other differentiae)
GENUS-TERM-1 = 通信_NOUN
GENUS-TERM-2 = 回線_NOUN

In this case, there are two words for each entry. If the numbers of words in the genus terms are the same, the words match one-to-one, that is:

[active line:活動回線]
→ [[telecommunication:通信],[line:回線]]
→ [line:回線]

This is knowledge that represents bilingual and disambiguated is-a relationships between the entry word and its genus term. The relationships constitute a disambiguated thesaurus. If the numbers of words in the genus terms are different, the alignment procedure is required, which is described in Section 3.

2.2 Extracting a Disambiguated Thesaurus by Using Parallel Corpora

One of the issues in acquisition of is-a relationships from monolingual dictionaries is that words in the definitions contain ambiguities. Therefore, the words in the relationships must be disambiguated.

Use of parallel dictionaries makes it possible to extract disambiguated is-a relationships. For example, the parallel definition of the entry word "card column" is:

card column: カード欄
(E) a line of punch positions parallel to the shorter edge of a punch card.
(J) 穿孔カードの短い辺に平行な穿孔位置の行

From the definitions, the following relationships are extracted:

[card column:カード欄] → [line:行]

Both "card column" and "active line" have the genus term "line"; however, the meaning of "line" is different. The expression [line:行] represents the "line" in the definition means lines in images, while [line:回線] means electric lines. The granularity of word-sense is a serious problem affecting the acquisition of semantic knowledge. In this paper, a disambiguated word is presented by a translation pair such [line:行] or [line:回線]. The disambiguation level is useful when the knowledge is used for practical applications such as machine translation.

3 An Algorithm for the Extraction

3.1 Recognition of the Genus Terms of Entries

In Section 2, the acquisition of genus terms was described. Since the dictionary we used is for technical terms, the genus term often consists of multiple words. In this paper, the longest possible genus term for an entry is recognized. The genus term is a word sequence that contains the parts of speech of the entry, and also possibly adjectives, and adverbs. To absorb the differences between sets of parts of speech in English and Japanese, some modifications are needed:

- In English, the pattern "NOUN1 of NOUN2" in a genus term is transformed into "NOUN2 NOUN1." In Japanese, the pattern "NOUN1 の NOUN2" is transformed into "NOUN2 NOUN1".
- Adjectives and adverbs are treated as the same syntactic category.

The matching between words in genus terms is not always one-to-one. For example, suppose that the English genus terms "direct_ADJ addressing_VERB mode_NOUN" and the Japanese genus terms "直接_ADV アドレス_NOUN 指定_NOUN モード_NOUN" are aligned. The English pattern consists of three words, while the Japanese pattern consists of four words. For many-to-many matching, a method

for sentence alignment using dynamic programming framework developed by Gale et al. [5] is used. As they claim, the framework is useful when sequences such as sentences are compared by using a distance measure, which they calculate by using a probabilistic model. We use the framework to align the words in genus terms. As distance measure, we use the syntactic categories of the words and co-occurrence in the parallel dictionary and bilingual corpora.

4 A DP Algorithm for Genus Term Alignment

The algorithm for aligning words in the genus terms is basically the same as Gale's without the calculation of the distance measure, which we call the preference calculation.

Let ew(i) (i=0, ...) be the (i+1)th word in the English genus term, and let jw(j) (j=0, ...) be the (j+1)th word in the Japanese genus term. P(i,j) is the preference between the word sequences $ew(1), \ldots, ew(i)$ and $jw(1), \ldots, jw(j)$. Suppose that p is a preference function. For example, suppose that p(ew(1), jw(1); 1, 1) represents a match between ew(1) and jw(1) (one-to-one matching). P(i,j) is calculated according to the following formula:

P(i,j) = max(P1,P2,P3,P4,P5)
P1 = P(i-1,j-1) + p(ew(i-1),jw(j-1);1,1)
P2 = P(i-2,j-1) + p(ew(i-2),jw(j-1);2,1)
P3 = P(i-1,j-2) + p(ew(i-1),jw(j-2);1,2)
P4 = P(i-1,j) + p(ew(i-1),jw(j);1,0)
P5 = P(i,j-1) + p(ew(i),jw(i-1);0,1)

Suppose P(0,0) = 0. The preference function p reflects the following two factors:

- Alignment between words that have the same syntactic category is preferred.
- Alignment between words that co-occur in other sentences in the parallel dictionary or corpora is preferred.

The preference function p for alignment of k words from *jw(i)* and l words from *ew(j)* is calculated by using the following formula. The function *Syn_cat(w)* returns the syntactic category of the word w.

p(jw(i),ew(j);k,l) = category_preference(jw(i),ew(j),k,l) + co-occurrence_preference(jw(i),ew(j),k,l)

category_preference(jw(i),ew(j),k,l) =

$$\begin{cases} 1 & : (k = 1 \textit{ and } l = 1) \textit{ and } (syn\_cat(jw(i)) = syn\_cat(ew(j))) \\ 0.75 & : (k = 1 \textit{ and } l = 1) \textit{ and } (syn\_cat(jw(i)) \; != \; syn\_cat(ew(j))) \\ 0.5 & : k = 2 \textit{ or } l = 2 \\ 0 & : k = 0 \textit{ or } l = 0 \end{cases}$$

co-occurrence_preference(jw(i),ew(j),k,l) =

$$\begin{cases} \frac{n}{m} (m > 0) : \textit{Here, } n \textit{ is the number of definitions that contain the same alignment of words but do not contain the same differentiae, while } m \textit{ is the number of the total number of definitions that contain the same alignment of words.} \\ 0 : (m = 0) \end{cases}$$

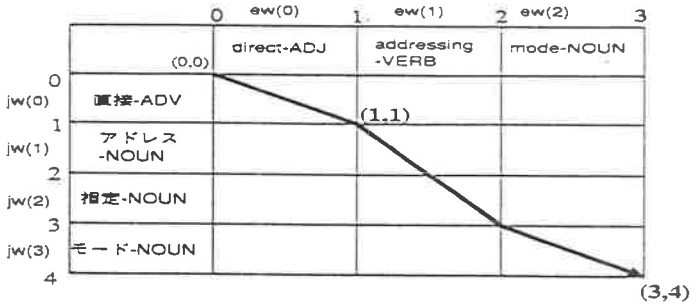

Figure 3: Alignment Matrix

For example, p(jw(1),ew(1);2,1) gives the preference for matching of one English word, "addressing," and two Japanese words, "アドレス" and "指定". The matching is one-to-two, so the category_preference is 0.5. For the co-occurrence_preference, the following bilingual definition is found among the definitions:

(E) `ACF/TCAM, any point-to-point line configuration in which the station on the line does not use polling and <<addressing>> characters.`
(J) ACM/TCAM において、回線上のステーションがポーリングと<<アドレス指定>>文字を使用しないポイントツーポイント回線構成

If the number of the definition that contains the words "addressing" is two, the preference is $\frac{1}{2}$. Therefore p(jw(1),ew(1);2,1) is $0.5 + 0.5 = 1.0$.

To align the words, an alignment matrix is created. Figure 3 shows the matrix for the example. Rows in the matrix show the sequence of English words, and columns represent the sequence of Japanese words. The position (i,j) in the matrix represents P(i,j). The path from (0,0) to (3,4) in the matrix represents the alignment of the words.

In this case, the shortest path is [(0,0)→(1,1)→(2,3)→(3,4)], which gives a correct alignment.

5 Experiments

5.1 Experiment-1: Extraction from a Parallel Dictionary

We concluded a small experiment on extraction of a disambiguated thesaurus of 1,000 pairs of definitions for entries that begin with "a". The matching of genus terms in definitions was done by a grep-like tool is called parallel grep (PGREP). As options, PGREP requires English and/or Japanese patterns and actions during pattern matching. If the matching of the genus term was not one-to-one, the words were aligned by dynamic programming. The results of the alignment were compared with a human's answers. We obtained a correct alignment rate of 91.3 % for the 1,000 sample definitions.

The main cause of failure was the difference in word order of English and Japanese genus terms. Another problem is that the pattern "WORD1 WORD2 of WORD3," which is very common in the definitions, is translated in various ways.

5.2 Experiment-2: Extraction from a Parallel Corpora

In experiment-1, we extracted is-a relationships from definitions that have a specific pattern. However, since there are few parallel dictionaries, we had to extract the relationships from "ordinary" bilingual

corpora. In our second experiment, we used bilingual (English and Japanese) computer manuals. Hearst proposes a method for extracting is-a relationships from a monolingual encyclopedia by using patterns "such NP as NP" and "NP, including NP" [8]. We use the simpler pattern "NOUN is a/the GENUS-TERM." The pattern cannot be used for monolingual text, since the many meanings of "be" cause ambiguities. However, by using both English and Japanese patterns, the number of ambiguities can be reduced, since some ambiguities in the sentences are resolved in their translation [17].

In the experiment, the following patterns are used (the expression "|" is an OR operator).

English pattern:
MOD* E-NOUN+ [is a|is the|are] MOD* E-GENUS-TERM*

Japanese pattern:
MOD* J-NOUN+ [は | が | とは] MOD* J-GENUS-TERM+ [です。 | で、 | である。 |。 | を言います。]

By using the pattern bilingual pair [E-NOUN:J-NOUN] → [E-GENUS-TERM:J-GENUS-TERM] was extracted.

For example, from the following sentences in the manual, the relationship [offset:オフセット] → [number:数] is extracted.

(E) If the <offset$_{E-N}$> *is a* negative <<number $_{E-GT}$>>, the routine associated with the offset probably did not allocate a save area, or the routine may have been called using SVC-assisted linkage.
(J) <オフセット$_{J-N}$>が負の<<数$_{J-GT}$>>である場合には、このオフセットに関連づけられるルーチンが保管域を割り振らなかったか、そのルーチンがＳＶＣ援助連係を使用して呼び出された可能性があります。

These patterns were applied to a bilingual manual text containing 2,000 pairs of sentences. Thirty-four pairs were matched of which 30 were correct. Most failures were caused by free translation.

6 Related Work

For the practical use of the alignment program, the following four issues concerning its accuracy must be considered.

1. Robustness when used with very large corpora
2. Use of practical computational resources
3. Language-independence of the algorithm
4. Use of information solely in the corpora to be aligned.

Since our approach uses simple pattern matching and dynamic programming for some words in sentences, the first two obstacles can be avoided. In our framework, information on the syntactic categories and co-occurrence of the words in the corpora is used. Though the set of syntactic categories needed depends on the language, our method can be applied if some heuristics are used.

In terms of knowledge acquisition from dictionaries, use of a parallel dictionary makes it possible to construct disambiguated (bilingual) relationships. One of the obstacles in the monolingual approach is ambiguity of words in definitions. Guthrie et al. proposed a method for extracting disambiguated is-a relationships from the LDOCE with matching of case-patterns [7]. However, since the number of semantic markers used in the LDOCE is relatively small, it is difficult to resolve ambiguities completely.

Our method uses bilingual definitions, so the word-senses of the words in the thesaurus are represented by pairs of English word and its Japanese translation. The notation is still ambiguous when the words in both languages contain the same ambiguities. However, the notation is useful when the knowledge is used in practical applications such as machine translation.

There have been some studies of the alignment of sentences [2, 6, 10]. These studies are classified according to what kind of "anchor point" binds both (parts of) sentences. Kay et al. use words appearing in the sentences [10]. Gale et al. use the lengths of sentence pairs [6], while Brown et al. employ the numbers of words in them [2]. In word-to-word alignment, the fact that the word order depends on the language prevents the development of a word-to-word alignment algorithm. In this paper, alignment of genus words that form a part of sentence is proposed. The framework developed for sentence alignment can be applied to our work, using the "genus term and differentia" structure in the definition sentences.

7 Conclusion

We have described a method for extracting disambiguated (bilingual) is-a relationships from parallel (English and Japanese) dictionary definitions by using word-level alignment: We are now evaluating the method by using more examples. Alignment of English and Japanese is more difficult than that of English and, say, French, since Japanese phrases have no word boundaries. Therefore, the recognition of compound nouns depends on the lexicon and the algorithm in the tagger. For example, it is clear that "telecommunication line" consists of two words. However, its Japanese translation "通信回線" can be recognized as one word or two words ("通信" and "回線"). To absorb the differences between the languages, a more refined algorithm is required.

In the work described here, only genus terms were extracted. However, definitions contain other useful information. The alignment of other parts of sentences is important. Knowledge extraction from an unrestricted bilingual corpus rather than from definition sentences is another challenging issue. Tools for dealing with bilingual data are also needed.

References

[1] R. A. Amsler. "A Taxonomy for English Nouns and Verbs". In *Proceedings of the 19th Annual Meeting of the ACL*, pages 133–138, 1981.

[2] P. F. Brown, J. C. Lai, and R. L. Mercer. "Aligning Sentences in Parallel Corpora". In *Proceedings of the 29th Annual Meeting of ACL*, pages 169–176, 1991.

[3] IBM Corporation. *IBM Dictionary of Computing (in Japanese)*, volume N:SC20-1699-07. IBM Corporation, 1991.

[4] I. Dagan, A. Itai, and U. Schwall. "Two Languages are More Informative Than One". In *Proceedings of ACL-91*, 1991.

[5] W. Gale and K. W. Church. "A Program for Aligning Sentences in Bilingual Corpora". In *Proceedings of ACL-91*, pages 177–184, 1991.

[6] W. A. Gale and K. W. Church. "A Program for Aligning Sentences in Bilingual Corpora". *Computational Linguistics*, 19(1):75–102, 1993.

[7] L. Guthrie, B. M. Slator, Y. Wilks, and R. Bruce. "Is There Content in Empty Heads?". In *Proceedings of COLING-90*, pages 138–143, 1990.

[8] M. A. Hearst. "Automatic Acquisition of Hyponyms from Large Text Corpora". In *Proceedings of COLING-92*, pages 539–545, 1992.

[9] IBM Corporation. *IBM Dictionary of Computing*, volume SC20-1699-07. IBM Corporation, 1988.

[10] M. Kay and M Roscheisen. "Text-Translation Alignment". *Computational Linguistics*, 19(1):121–142, 1993.

[11] J. Klavans, M. S. Chodorow, and N. Wacholder. "From Dictionary to Knowledge Base via Taxonomy ". In *Proceedings of the 6th Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary and Text Research*, pages 110–127, 1990.

[12] H. Maruyama and S. Ogino. "The Mega-Word Tagged Corpus Project". In *Proceedings of TMI-93*, 1993.

[13] M. McCord. "The Slot Grammar System". Technical Report RC17313, IBM Research Report, 1991.

[14] J. Nakamura and M. Nagao. "Extraction of Semantic Information from an Ordinary English Dictionary and Its Evaluation". In *Proceedings of COLING-88*, pages 459–464, 1988.

[15] P. Procter. *"Longman Dictionary of Contemporary English"*. Longman Group Limited, Harlow and London, England, 1978.

[16] N. Uramoto. "Lexical and Structural Disambiguation Using an Example-Base". In *The 2nd Japan-Australia Joint Symposium on Natural Language Processing*, pages 150–160, 1991.

[17] T. Utsuro, Y. Matsumoto, and M. Nagao. "Lexical Knowledge Acquisition from Bilingual Corpora". In *Proceedings of COLING-92*, pages 581–588, 1992.

[18] Y. Wilks, D. Fass, C. Guo, J. McDonald, T. Plate, and B. Slator. "Providing Machine Tractable Dictionary Tools". *Machine Translation*, 5:99–154, 1990.