기계번역에서 동사 모호성 해결에 관한 하이브리드 기법

문 유 진[†] · 마르타 파머[††]

요 약

본 논문에서는 기계번역에서 동사 번역의 모호성 해결을 위한 하이브리드 기법을 제안한다. 제안된 기법은 동사 번역을 위해 개념기반의 기법과 통계기반의 기법을 병행하여 수행하는 알고리즘이다. 이를 위해 연어사전, WordNet과 말뭉치에서 추출한 통계 정보를 이용한다. 동사 번역의 모호성을 해결하기 위하여 이 알고리즘은 기계번역의 트랜스퍼 단계에서 번역할 동사의 번역어를 찾는다. 그러나 만일 적절한 번역어를 찾지 못하게 되면, WordNet을 참조하여 번역 문장에서 동사의 논리적 제약어와 연어사전의 논리적 제약어들 사이의 단어간 유사도를 측정하여 번역어를 찾는다. 그리고 이와 동시에 이 알고리즘은 말뭉치에서 추출한 통계 정보를 참조하여 공기 유사도를 측정하여 번역어를 찾는다. 실험 결과, 이 알고리즘은 번역 정확성에서 기존의 다른 알고리즘보다 우수하며, 특히 연어기반의 기법과 비교할 때 약 24.8% 정도의 번역 정확성이 향상된 것으로 나타나고 있다.

A Hybrid Method of Verb Disambiguation in Machine Translation

Yoo-Jin Moon[†] · Martha Palmer[††]

ABSTRACT

The paper presents a hybrid method for disambiguation of the verb meaning in the machine translation. The presented verb translation algorithm is to perform the concept-based method and the statistics-based method simultaneously. It uses a collocation dictionary, WordNet and the statistical information extracted from corpus. In the transfer phase of the machine translation, it tries to find the target word of the source verb. If it fails, it refers to WordNet to try to find it by calculating word similarities between the logical constraints of the source sentence and those in the collocation dictionary. At the same time, it refers to the statistical information extracted from corpus to try to find it by calculating co-occurrence similarity knowledge. The experimental result shows that the algorithm performs more accurate verb translation than the other algorithms and improves accuracy of the verb translation by 24.8% compared to the collocation-based method.

※이 논문은 한국과학재단의 1996년도 후반기 해외 Post-Doc. 지원비에 의하여 연구되었음.
† 정 회 원 : 호남대학교 컴퓨터공학과
†† 비 회 원 : 펜실베니아 대학교 인지과학연구소
논문접수 : 1997년 9월 23일, 심사완료 : 1998년 1월 31일

1. Introduction

1.1. General Remarks

The ambiguity of the word meaning is one of the most frequently occurring phenomena in the translation of the source language. There are mainly two kinds of ambiguity for the word meanings - the structural ambiguity and the lexical ambiguity ([4]). This paper deals with the lexical ambiguity in the machine translation. To resolve the lexical ambiguities various kinds of method have been used, which are a dictionary-based method, a semantic feature method, a collocation-based method, a statistics-based method, an example-based method and a neural network method etc.([1, 10, 11, 12]).

This paper tries to resolve the ambiguities for verb translation in the Korean-English Machine Translation System (KEMT System), which consists of an analysis phase, a transfer phase and a generation phase ([4]). The analysis phase consists of a lexical analysis phase, a syntactic analysis phase, a semantic analysis phase and a pragmatic analysis phase. The transfer phase consists of idiom to idiom translation, collocation translation and word to word translation. The generation phase generates the target language sentence by rearranging the output of the transfer phase using English grammar rules. The paper deals with collocation translation for verbs in the transfer phase. As well, the paper is consistent with directions in which research on language from both practical and theoretical perspectives appears to be evolving.

1.2 Statement of Problem

In a Korean-English dictionary, the Korean verb "타다(ta-da)" may be translated into many English meanings, i.e. dissolve, willow, play on, strum on, twang on, play, be under, be sensitive to, be easily damaged from lack of rain, walk, scale, get on, climb, ride in, take and ride, etc. ([6]). When "타다(ta-da)" in a Korean sentence is translated into an English phrase, the KEMT System should decide which English word should be taken for translation.

The verb translation problem in the KEMT System is a little different from the problem of word-sense disambiguation for a polysemous verb, which disambiguates a sense among the various senses of the polysemous verb. A sense for a verb selected from the various senses may be translated into multiple words in a target language. For example, the polysemous verb "타다(ta-da)" has a sense "ride". Korean sentences "기차(train)를 타다" and "자전거(bicycle)를 타다" are translated into "take a train" and "ride a bicycle" in the KEMT System, even though both belong to the same sense of "타다(ta-da)". In this paper, both are treated as different senses of "타다(ta-da)".

2. Literature Review for Korean Verb Translation

2.1. a Statistics-Based Method

Resnik ([9]) says that statistics-based methods in natural language processing are moving toward the integration of more linguistic information into probabilistic models - as an indication of how much the Penn Treebank is moving in the direction of annotating not only surface linguistic structure but predicate argument structure as well. This makes sense, since the value of a probabilistic model is ultimately constrained by how well its underlying structure matches the underlying structure of the phenomenon it is modeling.

[11] suggests the combined method of a collocation-based method and a statistics-based method. It calculates co-occurrence similarity knowledge between words using statistical information from corpus. Verbs are translated using the similarity match, when the verb-related nouns do not exactly match the collocations specified in the dictionary. It shows about 88% accuracy for the Korean verb translation. It works well in the specific domain for which knowledge of the co-occurrence similarity has been built. But statistical information for the general domain is

not sufficient, and it does not work well in the general domain.

2.2. a Concept-Based Method

[6] suggests a concept-based method for Korean verb translation, which is the combined method of a collocation-based method and an example-based method. The transfer phase in the KEMT System refers to the idiom dictionary to find translated English words for Korean verbs, and if it fails, it refers to the collocation dictionary to find them. If that fails, a concept-based verb translation is performed. The concept-based verb translation refers to the collocation dictionary once more to find the conceptually close sense of the input Korean verb, refers to WordNet to calculate word similarities among the input logical constraints and those in the collocation dictionary, and selects the translated verb sense with the maximum word similarities beyond the specified critical value.

It shows about 91% accuracy at the critical value 0. 4 when applied to the 5th grade student textbooks. It works well for the general domain, but not for the specific domain. It is a kind of the example-based method, using Korean WordNet which will be explained in section 3.1.4.

2.3. Word Similarity

Many factors influence judgements of semantic similarity between two nouns. A great many researchers ([2, 3, 8]) are investigating techniques for deriving measures of word similarity on the basis of distributional behavior. [9] opted to use taxonomic relationships in WordNet as the basis for an information-theoretic similarity measure. Like the formalization of selectional preference, this has the advantage of combining inductive, quantitative methods with an existing broad-coverage source of lexical knowledge. Counting links is to consider the information content of a class as a way to measure its specificity. Information content of a class is defined in the standard way as negative the log likelihood, or log 1/p(c). The simplest way to compute similarity of two classes using this value would be to find the superclass that maximizes information content; that is, to define a similarity measure as follows:

(1) $WS(c_1, c_2) = \max\{\log 1/p(c_i)\}$.

where $\{c_i\}$ is the set of classes dominating both c_1 and c_2, and the similarity is set to zero if that set is empty.

[7] says that word similarity from noun A to noun B in WordNet can be calculated by measuring how close common superordinates of the two nouns (A and B) are, which can be calculated by the expression (2) below.

(2) WS(A, B)

$$= \frac{(\#\text{ of common superordinates of A and B}) * 2}{(\#\text{ of superordinates of A and B})}$$

To calculate the word similarity from noun A to noun B using the expression (2) requires only hypernyms of nouns among various types of information for nouns in WordNet. It models human cognitive processes and effectively selects the similar words for the given word. Thus it will be used in this paper.

Until now, there is no agreement which measurement is the most accurate. The more optimized a measurement tries to be, the more complicated it becomes and the longer it requires the calculation time. And the quality of the word similarity measurement depends on whether it selects the similar words to the given word, not on the figures. Therefore, this paper uses the effective and simple measurement for word similarity measurement.

2.4. Co-occurrence Similarity

[11] classifies the set of relations G between a noun and a verb into five grammatical relations as follows.

$G \equiv \{sbj, obj, loca, inst, modi\}$

And he defines the set of co-occurrence verbs $V_g(n)$ for a noun n as follows. $f_g(n, v)$ is the co-occurrence frequency from corpus between a noun n and a verb v in the grammatical relation g.

$V_g(n) \equiv \{v | v$ is a verb such that $f_g(n, v) \geq 1\}$,
where $g \in G \equiv \{sbj, obj, loca, inst, modi\}$

The co-occurrence similarity $|V_g(n)|$ is the sum of the co-occurrence frequencies among a noun n and verbs v in the grammatical relation g.

$$|V_g(n)| \equiv \sum_{v \in V_g(n)} f_g(n, v)$$

The set of relations G for the co-occurrece similarity $|V_g(n)|$ may contain any other relations than the above described G. But the paper utilizes "sbj" and "obj" relations among the set of relations G for the co-occurrece similarity $|V_g(n)|$.

3. Statement of Methodology Used

3.1. Representation of Knowledge

3.1.1. The Korean-English Dictionary

The Korean-English dictionary contains Korean entries, and domain default and normal default of translated English words for each Korean entry. The domain default means default of translated words when domain is specified. The normal default means default of translated words in a general domain.

3.1.2. The Idiom Dictionary

The idiom dictionary for the KEMT System contains entries for Korean idioms, for domain idioms, and for normal idioms of translated English words for each Korean idiom.

3.1.3. The Collocation Dictionary

The collocation dictionary for the KEMT System has been built by extracting collocation lists from Korean-English dictionaries and corpus. It contains entries for the Korean collocations, domain collocations, and normal collocations of translated English words for each Korean collocation. In (Fig. 1) "타다(ta-da)" has four meanings - burn, dissolve, play on and climb. If the subject of a sentence is "tree" and the predicate is "타다(ta-da)", they are translated into "a tree burns". If the object is "tree" and the predicate is "타다(ta-da)", they are translated into "climb a tree".

```
"타다"   :6 normal collocation
      SUBJECT
        "burn"      "나무"=> tree
                    "머리카락"=> hair
                    "장작"=> wood

      OBJECT
        "dissolve"  "소금"=> salt
                    "물감"=> dye
        "play"    "거문고"=> keomungo
        "climb"     "나무"=> tree
                    "산"=> mountain
          .
          .
          .
```

(Fig. 1) Normal collocations in the collocation dictionary

3.1.4. Korean WordNet

WordNet for English was constructed by Miller, Beckwith, Fellbaum, Gross and Miller at Princeton University in 1990 ([5]). WordNet can be said to be a dictionary based on psycholinguistic principles.

Korean WordNet (KWN) are actually sets of Isa hierarchies of Korean nouns. The Isa hierarchies consist of nodes and edges. The nodes are represented by synonym sets of English WordNet and those of Korean nouns. And the edges are represented by hypernymous relations among nodes ([7]).

3.2 A Hybrid Algorithm for Disambiguation of Korean Verb Translation

The transfer phase in the KEMT System refers to

the idiom dictionary to find translated English words for Korean verbs, and if it fails, it refers to the collocation dictionary to find them. If that fails, the following hybrid algorithm is performed. The algorithm stages 2) and 3) can be performed simultaneously. Logical constraints are limitations on the applicability of predicates to arguments.

1) The transfer phase refers to domain collocation lists and normal collocation lists in the collocation dictionary in order to find the conceptually close sense of the input Korean verb.

2) It refers to KWN to calculate word similarities in sequence between the logical constraint of the input verb and that of the collocation list. (See section 2.3.)

It selects the translated verb sense with the maximum value of the word similarity beyond the critical value 0.4 ([6]).

3) if it can't select the translated verb sense in the above stage, it refers to statistical information to calculate co-occurrence similarities in sequence between the logical constraint of the input verb and that of the collocation lists. (See section 2.4.)

It selects the translated verb sense with the maximum value of the co-occurrence similarity beyond the critical value.

4) If the results of the stage 2) and 3) are null, go to the stage 5).

If the results of the stage 2) and 3) are the same, return the results.

If the result of the stage 2) is null, return the result of the stage 3).

If the result of the stage 3) is null, return the result of the stage 2).

5) If it can't select the translated verb sense in the above stages, it selects the default of the translated verb sense.

The logical constraint of the input verb means the object of the Korean input, if the logical constraints in the collocation dictionary belong to an object. Otherwise, it means the subject of the Korean input. As well, the logical constraint of the collocation list means the Korean object or subject of the corresponding one in the collocation list to the input verb.

The verb in the sentence "가야금을 타다" can be translated into "play kayakeum" in the stages 1) and 2). The algorithm uses the collocation dictionary (Fig. 1) and KWN for word similarities between "거문고 (keomungo)" and "가야금 (kayakeum)", which is a concept-based method. The verb in the sentence "계단을 타다" can be translated into "climb the stairs" in the stage 3) using the statistical information and corpus for co-occurrence similarities between "산 (mountain)" and "계단 (stairs)", which is a statistics-based method.

4. Results

The algorithm suggested in section 3. 2 has been applied to the KEMT System for the fifth-grade student textbooks (about 800KB) of the general domain and IBM manuals (about 1.2MB) of the computer science domain. It works well in the sentences of the general domain, whose verbs are mainly translated during the stage 2) in section 3.2. Also it works well in the sentences of the computer science domain, whose verbs are mainly translated during the stage 3) in section 3.2. While the collocation-based method shows about 69.8% of accuracy in the verb translation for the above mentioned domains, the hybrid algorithm as illustrated in 〈Table 1〉 shows about 94.6% accuracy and improves accuracy of the verb translation by 24.8% compared to the collocation-based method.

〈Table 1〉 Comparison of the Methods for Verb Translation

	Size of Texts	Accuracy of Verb Translation
The Collocation −Based Method	about 2M Bytes	69.8%
The Hybrid Method	about 2M Bytes	94.6%

〈Table 2〉 Success Ratio of the Verb Translation According to the Number of Words

No. of Words in a Sentence	No. of Sentences	Success Ratio
1 – 5	9,302	97.5%
6 – 10	12,104	91.1%
11 – 20	7,725	95.1%
Total	29,131	94.6%

And the hybrid method performs more accurate verb translation than the concept-based method and the statistics-based method. 〈Table 2〉 indicates that the number of words in a sentence has nothing to do with the success ratio of the verb translation and kinds of the verb have something to do with the success ratio.

In addition, experiments show that it heavily depends on the contents of the collocation dictionary. Incorrect translation of the input Korean verb occurs when the collocation dictionary contains collocation lists which have been incorrectly translated into English, when the collocation dictionary contains idiom lists which are fixed expressions, and when the input Korean verb or the logical constraints are not entries in a Korean-English dictionary. Therefore, gradual updates of the collocation dictionary are required.

The method suggested in the paper works for the input logical constraints with ambiguous Korean nouns. An input logical constraint with an ambiguous Korean noun may be translated into more than one sense, of which the closest sense with the maximum value of similarity to the logical constraints in the collocation dictionary is selected and the other senses are rejected. In reality, the correct sense of the ambiguous noun coincides with that of the collocation list and results in the sense with the maximum value of similarity, and therefore the ambiguous Korean nouns work quite well.

5. Conclusions

The paper presents a hybrid method for disambiguation of the verb meaning in the machine translation. The presented algorithm has been applied to the sentences of the general domain and those of the computer science domain. It works quite well in the general domain, as well as in the computer science domain. While the collocation-based method shows about 69.8% accuracy in the verb translation for the above mentioned domains, the hybrid algorithm shows about 94.6% accuracy and improves accuracy of the verb translation by 24.8% compared to the collocation-based method. And it performs more accurate verb translation than the other algorithms.

In addition, the presented method may be applied to the verb translation for other languages.

Future tasks to be completed are as follows.

First, KWN should be extended to Korean nouns and verbs and to each specific domain. Because existing KWN has been built only for the general domain of about 17,000 Korean nouns. Second, gradual updates of the collocation dictionary are required, since the algorithm heavily depends on the contents of the collocation dictionary.

References

[1] W.Dolan, "Word Sense Disambiguation: Clustering Related Senses", Proc. of COLING-94, pp. 712-768, Aug. 1994.

[2] P.Brown et al., "Class-Based n-Gram Models of Natural Language", Computational Linguistics, pp.467-480, Dec. 1992.

[3] M.Hearst and H.Schutze, "Customizing a lexicon to better suit a computational task", Proceedings of the ACL SIGLEX Workshop, Columbus, Ohio, June 1993.

[4] Y.Kim, "Natural Language Processing", pp. 22-40, Kyohaksa, 1994.

[5] G.Miller, R.Beckwith, C.Fellbaum, D.Gross and K.Miller, "Introduction to WordNet: an On-line Lexical Database", Report of WordNet, Princeton

Univ., 1993.

[6] Y.Moon and Y.Kim, "Concept-Based Verb Translation in the Korean-English Machine Translation System", Journal of Korea Information Science Society (KISS), pp.1166-1173, Vol.22, No.8, Aug. 1995.

[7] Y.Moon, "Design and Implementation of WordNet for Korean Nouns Based on the Semantic Word Concept", Ph.D. Thesis, Dept. of Computer Engineering in Seoul National Univ., Feb. 1996.

[8] F.Pereira, N.Tishby and L.Lee, "Distributed clustering of English Words", Proceedings of ACL-93, June 1993.

[9] P.Resnik, Selection and Information: A Class-Based Approach to Lexical Relationships, Ph.D. Thesis, Univ. of Pennsylvania, pp.105-114, 1993.

[10] N.Uramoto, "Disambiguation with Distinctive Features Extracted from an Example-Base", NLPRS, pp.44-50, Dec. 1993.

[11] J.Yang, "Co-occurrence Similarity of Nouns for Ambiguity Resolution in Analyzing Korean Language", Ph.D. Thesis, Dept. of Computer Engineering in Seoul National Univ., Feb. 1995.

[12] D.Yarowsky, "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", Proc. of COLING-92, pp. 454-460, Aug. 1992.

문 유 진

1979년 한국외국어대학교 졸업.
1986년 펜실베니아 주립대학교 전산학과 졸업(이학석사).
1996년 서울대학교 컴퓨터공학과 졸업(공학박사).
1986년~1988년 삼성HP 근무.
1996년~1997 펜실베니아대학교 인지과학연구소 Post-Doc.
1991년~현재 호남대학교 컴퓨터공학과 조교수.
관심분야: 자연언어처리, 인공지능, 데이타베이스임.

마르타 파머

University of Texas, Austin 전산학과 졸업 (학사, 석사).
Edinbugh University 전산학과 졸업 (박사).
Unisys 근무.
University of Pennsylvania 전산학과 연구 부교수 및 인지과학연구소의 프로젝트연구개발부장임.
관심분야: machine translation, lexical semantics, knowledge base임.