Automatic Sense Disambiguation for Target Word Selection

Kwon Yang Kim
Dept.of Computer Science
Kyungpook Sanup University
Hayang, Korea

Se Young Park
Computer Dept.
ETRI
Taejon, Korea

Sang Jo Lee
Kyungpook
National University
Taegu, Korea

Abstract

This paper describes a method of automatic sense disambiguation for target word selection in Korean to English machine translation. At first, we define the concept of cluster for each sense of given verb according to corresponding target word. And then, we propose a method which selects the sense combination of words as the correct sense that has the greatest number of overlaps between input case slots and the predefined clusters for the given verb.

1 Introduction

The problem of word sense disambiguation is one which has received increased attention in recent work on natural language processing applications such as machine translation and information retrival. Given an occurrence of a polysemous word in text, it is needed to examine a set of senses in the dictionary and detect the clues for deciding intended sense in the context. Recently, several researches have been experimented with target word selection based on example-based machine translation[1,2,3]. There is no failure of selecting target word because the method is to select most similar one in a thesaurus than exact correspondent. However, the similarity methods based on thesaurus have some problem that it is difficult to make good thesaurus which provides accurate distance measure between words.

In order to solve the imperfection problems of disambiguation methods based on similarity, some researches using the exact matching with collocation extracted from bilingual corpus were experimented by [4,5]. But, these methods have a drawback that they can not select a target word when input sentences have nouns which were not included in collocation list because they need exact matching of collocation. And it is not clear how large the bilingual corpus would have to be become authoritative for disambiguation as an application independent collocation.

The well-known attempts to utilize information in MRD(Machine Readable Dictionaries) for lexical disambiguation are that of [6,7,8] which select the correct sense of a word by counting the overlaps between a dictionary sense definition and the definition of the nearby in the phrase. The sense of a word with greatest number of overlaps with senses of other words in the sentence is chosen as the correct one. These methods based on a small number of overlap between sense definitions is weak as a clue for disambiguation. As a result, these methods could not have certainty about disambiguation, because the relation is wholly dependent on word usage in the sense definition of a particular dictionary. Another variations of this idea are based on co-occurrence statistics, which is meant the preference two words appear together in the same context [9,10,11]. But, their method has same problems of MRD approaches.

Our method is based on the idea that it selects the sense combinations of words as the correct sense with the greatest number of overlaps between input case slots and the predefined clusters which are grouped according to corresponding target word.

2 Target Word Selection for Polysemy

The verb "쓰다" is a typical Korean polysemy. This verb has about ten sub-entries in a Amun Kak Korean dictionary and has twenty one translatable English verbs such as "write", "spend", "wear" and "adopt" in a Sisa Korean-English dictionary. Moreover, nouns of case slots which affect for selecting exact sense for the given verb have several senses. Therefore it is not easy to analyze the exact meaning and select the target word from sentences including ambiguous words.

When a given verb has a specific sense, words group of some case slots which occur with the verb may be thought of as a clue for disambiguating senses. We define the concept of cluster as a group of words which co-occur with a specific sense of given verb. Assume that we know of four senses of the verb "쓰다" and the nouns of case slots OBJ for each sense of the verb.

wear={모자/hat,안경/spectacles,가면/mask,...} adopt={방법/method,수단/ means,...}
write={소설/story,수필/ essay,책/ book,...}   spend={시간/time,돈/money,경비/expense,...}

Our method decides the correct target word by calculating the similarity between nouns of input sentence and nouns of predefined cluster for a specific sense of given verb. The similarity is based on overlap counts between definition of input nouns and common definition of predefined cluster. We assumed that the words within the cluster are similar each other and have the common features in the definitions of each words. Therefore, if the overlap counts between the sense definition of input noun (OBJ) and the definition of "wear"/OBJ cluster are greater than of "write"/OBJ, spend/OBJ and adopt/OBJ cluster, we can expect the sense of a given verb "쓰다" to "wear". Moreover, the another case slots "에" (wear/on, write/for, write/on) or "로" (write/with, write/in) will provide an additional clue for disambiguating them in addition to OBJ case slot.

This is similar to MRD-based attempts in the view of counting the overlap. However, there are distinct differences that we consider the overlap between the definition of input case slots and the predefined definition of the clustering which were grouped by the difference of correspondent target words rather than the definitions of the nearby in the phrase.

3 Clustering and Normalization

If a word has several senses, then the word is associated with several different set of co-occurring words, each of which corresponds to one of the senses of the words. In general situation, polysemous verb co-occurs with a large groups of nouns and one has to divide the group of nouns into a set of subgroups each of which correctly characterized the context for a specific sense of the polysemous word.

If we can determine which words co-occur with each word sense for a given verb, we can use these words for disambiguating the word in a given sentence. A cluster is defined as a group of nouns in each case slots of a given verb which are divided by corresponding English verbs and case slots. We used the four clues in order to divide the words group into the clusters for the given verb.

[Clue 1] Difference of corresponding English verbs: When one Korean verb has multiple senses, each senses are translated into different English verb according to some case slots for a given verb. The correct sense of transitive verb generally depends on object case slot.
[Clue 2] Difference of corresponding English cases(preposition): When one Korean verb has multiple senses, some case slots with same case marker are translated into different English preposition.
[Clue 3] Difference of classification number of word meaning in a Korean/English dictionary: Although the target verbs are same, it should be considered as different cluster if the classification number of word meaning is different in a Korean/English dictionary.
[Clue 4] Equivalence of target verbs: Although the target verbs are different, it should be considered as same cluster if the classification number of word meaning is same in a Korean/English dictionary.

The decision on corresponding target words is based on the usages of a given verb in Sisa Korean English dictionary and the results are reviewed by the Compton's Interactive Encyclopedia and a English native speaker.

For normalization of nouns in the cluster, we considered the features of nouns as the sense definition of the noun in the Gemong Korean encyclopedia. All the common function words such as case marker, determiner, numeral and conjunction should be removed in the definition. Function words tend to appear very often in the sense definitions for syntactic and style reasons rather than pure semantics. The words in definition are stemmed in order to match the derived or inflected form of same word together.

The normalization of nouns in the cluster is to extract a set of words that were occurred frequently in the definions of the each nouns in the cluster. The normalization process is divided into three sub-processes:

1. Union the definitions of nouns in cluster
2. Sort the definition words in cluster by relatedness
3. Extract the words group over a specific threshold

The relatedness represents the degree of importance of each words in the definition words of entry word. We used the following function which is assymmetric, so that relatedness(x,y) may not be equal to relatedness(y,x).

relatedness(x,y) = occur(x,y)/length(x)

We denote by occur(x,y) the occurrence of word y in a definition of entry word x and length(x) the total number of definition words in definition of entry word x. This relatedness function is different to the one used by [11]. We choose the some threshold values for the normalization of each clusters in our experiments. Followings are normalization results of "wear" cluster (verb "쓰다") for object case slot. This is a normalization result of 32 nouns in "wear/obj" cluster by the threshold 50.

dict(wear/obj,[쓰다/wear.만들다/make.머리/head.모직/woolen_fabric.쓰개/ hood.모양/shape, 관/crown,사용/use,가면/mask.꾸미다/decorate,재료/material.안경/spectacles.렌즈/lens, 달다/hang.족두리/headpiece.장식/ornament,위/on,벙거지/headgear,막다/cover,크다/big, 위하다/for,다리/leg.구슬/beads,가발/wig,투구/helmet.차림/attire,연극/play.여자/woman, 서양/western.보호/protection.머리털/hair,망건/headband.넓다/wide.남자/man.길다/long, 갖추다/prepare,가리다/cover.환관/coronet,중국/China.전립/felt_hat,입다/wear,얼굴/face, 악귀/demon.상모/hood.사람/person.부녀자/woman,갓/kat,놀이/play.털/hair.천/fabrics]).

4 Scoring Mechanism

We introduce the scoring mechanism for selecting target word. We define it based on the overlap counts between predefined cluster and input case slot. The scoring method finds nouns in case slots of input sentence and extracts the possible senses definitions of each nouns. The program generates the sense combinations by going through the sense definitions for noun in each case slot one by one. The scoring for these sense combinations is given by taking the defintions between input nouns and cluster of the each case slot and counting the overlap of definition words between them. For each combination, the total score is the sum of all the overlaps. This means that the program counts the overlap for the number of case slots which affect the deciding of corresponding target word in input sentence and sum them together.

score of each sense combinations =

$$\sum_{i=1}^{n} \text{overlap}(IC_i^j, CC_i) = \text{overlap}(IC_1^j, CC_1) + \text{overlap}(IC_2^j, CC_2) + \ldots + \text{overlap}(IC_n^j, CC_n)$$

In the score, IC_i^j means jth definition of input noun in case slot_i , CC_i means cluster defintion of case slot_i for a given verb and n means number of case slots which affect to disambiguation. Our scoring method is different to that of [6,7,8]. Lesk[6] simply counted overlaps by comparing each sense definition of a word with all the sense definition of the other words. Guthrie [7] use a similar method except that they put subject code as a single word in the definition list. Demertriou's scoring method [8]counts the sum of all the overlaps pairwise for each sense combination. So if the sentence has n ambiguous words, the program counts the overlap for all n!/(2!(n-2)!) pair combinations and add them together. Our method counts overlaps by comparing the definition of input case slots with predefined definitions of cluster for a given verb. This means that our method counts the overlap for the number of only predefined clusters in input sentence.

5 Experiments and Conclusions

To prove the effectiveness of our new mechanism, we extracted a number of sentences containing 5 Korean transitive verbs extracted from words of description in the newspaper Dong Ah. First, we considered the only object case slot with the given verb for selsecting target verb. The translation results are compared with human expert's intuition. In Table 1, results of experiments are shown. In the case of verb "쓰다", 54 of the 77 sentences were assigned sense correctly, giving a success rate of 70% and average overlap counts of 16 for the high score under the threshold 120. Of the 23 failures, 14 cases were used with case slots such as "에", "로" in the input sentences in addition to object case slot. Second, we did experiment on overlap with these case slots and sum them with the result of first experiment.

Of the 14 cases, 10 were assigned correct senses, giving an overall success rate of 83%. This indicates that "에" and "로" case slots are also clues for disambiguating in addition to object case slot. Although accuracy of the experiment results is 83%, our method confirms the potential contribution of the use of dictionary definition to the problem of lexical sense disambiguation for the Korean to English machine translation.

Verbs	No. of target word	Success % (Threshold)				
		180	150	120	90	60
verb1(쓰다)	18	69	69	70	70	60
verb2(대다)	16	67	68	68	65	59
verb3(타다)	15	70	71	69	64	60
verb4(넣다)	7	70	72	69	68	62
verb5(막다)	6	73	71	68	66	59

Table 1: Results of Experiments

References

[1] Sato, S. and Nagao, M. Toward Memory Based Translation, Proc. of COLING-90, p.247-252, 1990

[2] Kitano, H. A Comprehensive and Practical Model of Memory-Based Machine-Translation, p.1276-1282, IJCAI93, 1993

[3] Muraki, K. and Doi, S. Robust Translation and Meaning Interpretation Mechanism Based on Examples in Dictionary, PRICAI-92, 1992

[4] Lee, Ho S. and Kim, Yung T. The Collocation Structure for Disambiguation in English-Korean Machine Translation, p.213-218, PRICAI-92, 1992

[5] Ok, Cheol Y. A Selection of Best Translation Using Collocation, PRICAI-92, 1992

[6] Lesk, M. Automatic Sense Disambiguation Using Machine Readable Dictionaries:how to tell a pine cone from an ice cream cone, In Proceedings of the ACM SIG-DOC Conference, 1986

[7] Guthrie, Joe A., Guthrie, L. and Cowie, J. Lexical Disambiguation Using Simulated Annealing, In Proceedings of the 14th Conference on Computational Linguistics, Proc. of COLING-92, p.359-364, 1992

[8] Demetriou, George C. Lexical Disambiguation Using Constraint Handling In Prolog (CHIP), MSc Dissertation, School of Computer Studies, University of Leeds, 1992

[9] Wilks,Y., Fass, D., Guo, C. M., McDonald, J., Plate, T. and Slator, B. A Tractable Machine Dictionary as Resource for Computational Semantics, In B, Computational Lexicography for Natural Language Processing, Longman, 1989

[10] McDonald, J.E., Plate, T. and Schvaneveldt, R. Using Pathfinder to Extract Semantic Information from Text, In Pathfinder associate networks, 1990

[11] Guthrie, Joe A., Guthrie, L., Wilks, Y. and Aidinejad, H. Subject-Dependent Co-occurrence and Word sense Disambiguation, In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, p.146-152, 1991