Serdica J. Computing **6** (2012), 401–408

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

NOUN SENSE DISAMBIGUATION USING CO-OCCURRENCE RELATION IN MACHINE TRANSLATION

Changil Choe, Hyonil Kim

Abstract. Word Sense Disambiguation, the process of identifying the meaning of a word in a sentence when the word has multiple meanings, is a critical problem of machine translation. It is generally very difficult to select the correct meaning of a word in a sentence, especially when the syntactical difference between the source and target language is big, e.g., English-Korean machine translation. To achieve a high level of accuracy of noun sense selection in machine translation, we introduced a statistical method based on co-occurrence relation of words in sentences and applied it to the English-Korean machine translator RyongNamSan.

1. Introduction. In machine translation, WSD (Word Sense Disambiguation) is one of the most difficult problems and numerous works have been devoted to solving this problem. These works can be found in various repositories for machine translation, and some of them are listed at the end of paper

ACM Computing Classification System (1998): I.2.7.

Key words: Machine Translation, Statistical Method, Word Sense Disambiguation, Co-occurrence relation, Bilingual corpus.

[1–12]. Well known disambiguation methods, such as knowledge based, example based, and co-occurrence relation based method, as well as methods using linear programming and logic programming, have their own respective.advantages and disadvantages.

To solve the disambiguation problem of noun sense selection in machine translation we introduced a statistical method based on co-occurrence relation of words in sentences. Training data are automatically built from an English-Korean bilingual corpus, and an extended sense set is introduced to enhance efficiency of training. In section 2 and 3, we describe the details of our method. The analysis of the statistical experiment in subsection 4 shows that our approach is an effective method for WSD.

Statistical Method for Selecting Noun Sense using Co-occurrence Relation. In sentences, words have direct or indirect semantic relation with other words. That is, the words which frequently and simultaneously occur in same sentences are strongly semantically related to each other much, while the words which never occur simultaneously in sentences have no semantic relation. This relation between words, the so-called co-occurrence relation, plays an important role in WSD.

We call the set of words which have co-occurrence relation with word w its CWS (Co-occurrence Word Set) and denote by $C(w)$. The CWS of each word is extracted from an English raw corpus by applying statistical testing. The CWS of prepositions, articles, determiners and conjunctions are not considered.

Let $M(w)$ denote the sense set of a word w. The problem of estimating the optimal sense m^* of the word w in sentence s can be formalized as follows.

$$m^* = \arg\max_{m \in M(w)} P(w, m|s) \tag{1}$$

In a sentence, the words that have no co-occurrence relation with w do not affect the sense estimation of$w$. Therefore, we can rewrite (1) as follows.

$$\begin{aligned} m^* &= \arg\max_{m \in M(w)} P(w, m|C(w)) \\ &= \arg\max_{m \in M(w)} P(w, m) \cdot P(C(w)|w, m) \\ &= \arg\max_{m \in M(w)} P(m|w) \cdot P(w) \cdot P(C(w)|w, m) \\ &= \arg\max_{m \in M(w)} P(m|w) \cdot \prod_{c \in C(w)} P(c|w, m) \end{aligned} \tag{2}$$

It's not easy to estimate correctly the parameter $P(c|w,m)$ in (3), as we are faced with a data sparse problem in estimation.

To solve this problem, we consider the hypernym of each word, which has an is-a relation with the word in a hierachical sense structure such as Thesaurus or WordNet, and also consider the sense assigned to the word. We define the ESS (Extended Sense Set) $E(w)$ of sense set $M(w)$ as follows.

$$E(w) = M(w) \cup \{m' | isa(m, m') = true,\ m \in M(w)\}$$

Here $isa(m, m') = true$ means that sense m' is the hypernym of sense m. Substituting $M(w)$ of (2) with $E(w)$, we get the following expression:

$$\hat{m} = \arg\max_{m \in E(w)} P(m|w) \cdot \prod_{c \in C(w)} P(c|w,m) \tag{3}$$

From the definition of the ESS, we can estimate m^* by analyzing the solution $\hat{m}$ of (3).

Parameter Estimation. To estimate two parameters $P(m|w)$ and $P(c|w,m)$, a sense tagged English corpus. However, it is very expensive to build such large training data. We extract training data using E-K automatic alignment from an E-K bilingual corpus and dictionary. Figure 1 shows the parameter estimation process.

The estimation of parameters $P(m|w)$ and $P(c|w,m)$ from the training data is formalized as follows.

$$P(m|w) = \frac{\Sigma'_m CT(w, m') * WG(m, m') + 1}{CT(w) + N_1} \tag{4}$$

$$P(c|w,m) = \frac{\Sigma'_m CT(w, m', c) * WG(m, m') + 1}{\Sigma'_m CT(w, m') * WG(m, m') + N_2} \tag{5}$$

Here, $CT(w)$ is the count of times that w is semantically tagged, $CT(w, m')$ is the count of times that w is semantically tagged as m' and $CT(w, m', c)$ is the count of sentences in which the word assigned semantic tag m' and the word c appear simultaneously. $WG(m, m')$ represents the weight that means the possibility of replacing the semantic tag m with virtual semantic tag m'. We set this weight to be 1 in case that m is equal to m', to be between 0 and 1 in case that m' is the hypernym of m, and to be 0 otherwise. N_1 and N_2 are the constants for the smoothing.

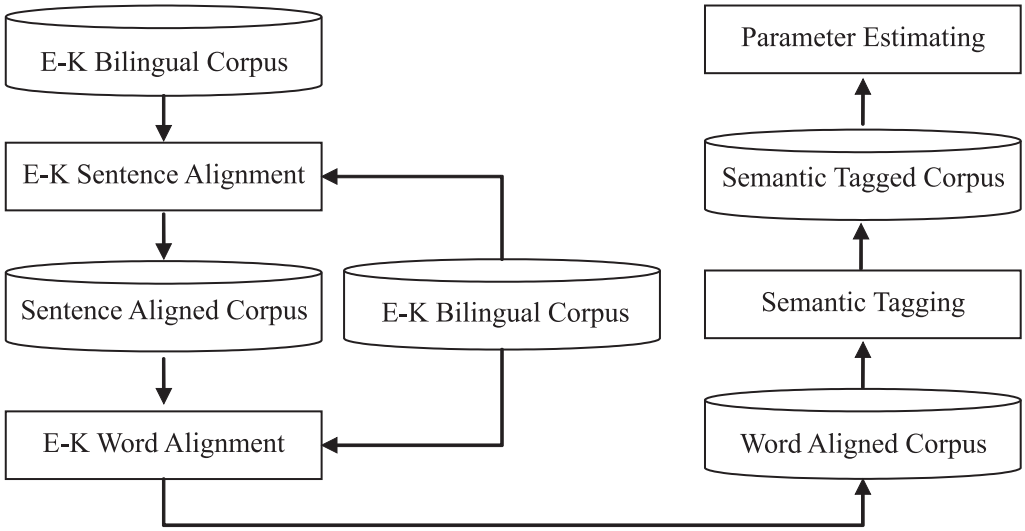

Fig. 1. Parameter estimation process

Experiments. We used the two polysemous words "plant" and "bank" to test the effect of our approach. Table 1 shows the sense set and ESS of the words "plant" and "bank" in the dictionary of the E-K machine translation system RyongNamSan. Table 2 shows the CWSs of the words.

Table 1. The sense set and ESS of "plant" and "bank"

Word	Sense set	ESS
plant	factory, plant, equipment	factory, work area, establishment, place, plant, living thing, equipment, goods, thing, lifeless thing
bank	enterprise, embankment, shore	enterprise, organization, social collective, composition, abstract thing, embankment, flood control equipment, service, establishment, public establishment, establishment, place, concrete thing, shore, far and near, space

Table 3 shows the result of extracting the sentences which include the word "plant" or "bank" from a semantically tagged English corpus. We used 80 percent of the total data to estimate the parameters and 20 percent to test our approach.

Table 2. CWSs of "plant" and "bank"

Word	Set of co-occurrence words
plant	animal, soil, root, seed, transgenic, growth, gene, nutrient, crop, water, leaf, produce, tissue, power, heat production, food, cell, fruit,...
bank	loan, credit, financial, central, capital, risk, lending, deposit, fund, river, cod, inshore, offshore, boat, ...

Table 3. Analysis of sentences which include "plant" or "bank"

Word	Number of sentences	Number of sentences by sense			
		Sense	Num of sentences	Training data	Test data
plant	24951	factory	10.098	8.091	2.007
		plant	13.094	10.475	2.619
		equipment	1.759	1.395	364
bank	9393	enterprise	5.230	4.180	1.050
		embankment	1.294	1.031	263
		shore	2.869	2.303	566

We measured the recall, precision and F-Score for each sense as follows. (We put the same weight on recall and precision for the calculation of F-Score.)

$$recall = \frac{the\ number\ of\ correct\ estimations\ to\ be\ m}{the\ number\ of\ total\ words\ which\ tagged\ with\ m} \times 100 \tag{6}$$

$$precision = \frac{the\ number\ of\ correct\ estimations\ to\ be\ m}{the\ number\ of\ estimations\ to\ be\ m} \times 100 \tag{7}$$

$$F - Score = \frac{2 \cdot recall \cdot precision}{recall + precision} \tag{8}$$

To measure the effect of the extended sense set, we performe two experiments. The first experiment is to measure the performance of the system without extended sense set, and the second experiment is to measure it with an extended sense set. The results are shown in Table 4 and Table 5.

Based on these results we calculated the recall, precision, and F-Score. In the first experiment average recall was 56.9%, average precision was 98.4% and F-

Table 4. The result of the first experiment

word	sense	Number of sentences	Num. of correct/num of total estimations	Recall [%]	Precision [%]	F-Score
plant	factory	2.007	987/1,004	49.2	98.3	65.6
	flant	2.619	1.741/1.763	66.5	98.8	79.5
	equipment	364	52/54	14.3	96.3	24.9
bank	enterprise	1.050	825/838	78.6	98.4	87.4
	embankment	263	12/12	4.6	100	8.73
	shore	566	293/301	51.8	97.3	67.6

Table 5. The result of the second experiment

word	sense	Number of sentences	Num. of correct/num of total estimations	Recall [%]	Precision [%]	F-Score
plant	factory	2.007	1.908/1.995	95.1	95.6	95.4
	plant	2.619	2.541/2,664	97.0	95.4	96.2
	equipment	364	329/331	90.4	99.4	94.7
bank	enterprise	1.050	1.031/1.067	98.2	96.6	97.4
	embankment	263	249/250	94.7	99.6	97.1
	shore	566	542/562	95.8	96.4	96.1

Score was 72.1, and in the second one average recall was 96.1%, average precision was 96.1% and F-Score was 96.1. These experiments show that introducing the extended sense set results in significant enhancement of average recall from 56.9% to 96.1%, but average precision fall downs from 98.4% to 96.1%. However F-Score is increased from 72.1 to 96.1.

Conclusion. In machine translation, semantic analysis is a very important process, but it is still very difficult because of the high cost of building a database. We introduced a method for noun sense disambiguation by building large amount of training data with low cost and using co-occurrence relation of

words. We could build training data easily from E-K bilingual corpus by realizing automatic E-K sentence alignment and word corresponding. And by introducing ESS using Thesaurus or WordNet, we could enhance training efficiency. Using ESS, we achieved significant enhancement of recall as 39.2%. The precision of our approach is measured as 96.08%.

REFERENCES

[1] Montoya A., A. Suarez, G. Rigau, M. Palomar. Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods. *Journal of Artificial Intelligence Research*, **23** (2005), 299–330.

[2] Fernandez-Amoros D., R. H. Gil, J. A. C. Somolinos, C. C. Somolinos. Automatic Word Sense Disambiguation Using Co-occurrence and Hierarchical Information. LNCS, Vol. **6177**, Springer, 2010, 60–67.

[3] Sumita E., H. Lida. Experiments and Prospects of Example-based Machine Translation. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, California, 1991, ©1991 Association for Computational Linguistics, 185–192.

[4] Jarmasz. M, S. Szpakowicz. Roget's thesaurus and semantic similarity. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03), Borovets, Bulgaria, 2003, 212–219.

[5] Specia L., A. Srinivasan, G. Ramakrishnan, M. Das Volpe Nunes. Word Sense Disambiguation Using Inductive Logic Programming. Inductive Logic Programming, Springer-Verlag, Berlin, Heidelberg, 2007, 409–423.

[6] Patwardhan. S, S. Banerjee, T. Pedersen. Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, 2003.

[7] Rosario. B, M. Hearst, C. Fillmore. The descent of hierarchy and selection in relational semantics. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02), Philadelphia PA, 2002, 417–424.

[8] Tratz S., A. Sanfilippo, M. Gregory, A. Chappell, C. Posse, P. Whitney. PNNL: A Supervised Maximum Entropy Approach to Word Sense Disambiguation. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, 2007, 264–267.

[9] Panagiotopoulou V., I. Varlamis, I. Androutsopoulos, G. Tsatsaronis. Word Sense Disambiguation as an Integer Linear Programming Problem. LNCS, Vol. **7297**, Springer, 2012, 33–40.

[10] Yarowsky. D. Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Cambridge, 1995, 189–196.

[11] Wilks Y., M. Stevenson. Word Sense Disambiguation using Optimised Combinations of Knowledge Sources. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, Canada, 1398–1403.

[12] Guo Y., W. Che, Y. Hu, W. Zhang, T. Liu. HIT-IR-WSD. A WSD System for English Lexical Sample Task. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, 2007, 165–168.

Changil Choe
Faculty of Mathematics
Kim Il Sung University
D.P.R.K
e-mail: `mathcci@yahoo.com`

Hyonil Kim
College of Computer Science
Kim Il Sung University
D.P.R.K
e-mail: `hyonilkim@yahoo.com`

Received August 28, 2012
Final Accepted October 25, 2012