Word Sense Disambiguation and Human Intuition for Semantic Classification on Homonyms*

Dong-Sung Kim[1] and Jae-Woong Choe[1]
[1] Department of Linguistics, Korea University, Anam-Dong 5 Ga, Seongbuk-Gu, Seoul, Korea
{dsk202, jchoe}@korea.ac.kr

Abstract. This paper reports a psycholinguistic research for the human intuition on the sense classification. The goal of this research is to find a computational model that fits best with our experiments on human intuition. In this regard, we compare three different computational models; the Boolean model, the probabilistic model, and the probabilistic inference model. We first measured the values of each models found in the semantically annotated Sejong corpus. Then the experimental result was compared with the values in the initial measurements. Kappa statistics supports that this agreement experiment is homogeneously coincidental. The Pearson correlation coefficient test shows that the Boolean model is strongly correlated with the human intuition.

Keywords: Human Intuition, Homonyms, Computational Model, Psycholinguistics, Word Sense Disambiguation, Boolean Model, Probabilistic Model, Probabilistic Inference Model

1 Introduction

Word Sense Disambiguation (WSD) is the basis of the natural language processing and other AI-related problems (Ide and Vernois 1998). Much of WSD mainly appeals to efficient computational techniques or some interesting discord in human annotators' classifications. In previous studies, human intuitive computational model on the sense classification is explained on the basis of estimation or probabilistic model (Lapata and Lascarides 2003; Lapata and Brew 2004). However, few of WSD work, if any, has addressed the issue of both linguistic and computational mechanism of the human intuition.

This paper is a psycholinguistic research on the computational model of human sense classifications. The well-known Distributional hypothesis (Harris 1964) claims that a sense of an ambiguous word is dependent on collocations around it. In other words, an analysis on the collocations elucidates the sense of their keyword(s). This also includes either statistical or logical behaviors of collocations in the context. Computational model is to predict such behaviors of collocating words. We, in this paper, compare three different models, one from each of the three bases, namely, logic, estimations, and probability. The particular models that are chosen are the Boolean model, the probabilistic model, and the probabilistic inference model. Each model also consists of several different versions of WSD methods. Among them, we choose the Boolean search, the Maximum Likelihood Estimation (MLE), and the naïve Bayesian classifier. The reason we choose those kinds of WSD methods is that they have previously been used for WSD technique and sense classification model (Ide and Vernois 1998, Lapata and Lascarides 2003, Lapata and Brew 2004), and thus have been proven to be relevant.

We conducted an experiment to observe the human intuition which provides us with clues to choice of the computational model. We investigate the correlation between each model's prediction and the experimental result. Also, we compare the decision made by each model with the sense classification made by humans.

The rest of the paper is composed of three parts: Section 2 is about three computational models. Section 3 describes the design of the experiment. Section 4 deals with the discussions.

* This work was supported by the Second Brain Korea 21 Project.

2 Computational Models

2.1 Boolean Model

Boolean model is the system to use a Boolean query that connects words with the conjunction operator, "AND." This model checks a set of contexts with a query string which consists of the collocation and the Boolean conjunction operator. If the query looks into the collocations of A and B, the query can be formularized as $WORD_A \cap WORD_B$. The query succeeds if it meets the context(s) containing "A AND B." Similarly, WSD is analyzed as searching the context with a Boolean query on every collocation. This type of approach is found in Mohammad and Pedersen (2004). They use the Boolean search model to search for a linguistic feature of an ambiguous word or some combinations of those features of ambiguous words in the context.

2.2 Probabilistic Model

Consider the hypothetical context of "$\alpha\ldots\mathbf{X_A}\ldots\beta$," which the probabilistic model characterizes as the joint probabilistic distribution $P(\alpha,\mathbf{X_A},\beta)$ of three variables. The distribution induces the formula: $P(\alpha,\mathbf{X_A},\beta) = P(\alpha|\mathbf{X_A},\beta)\cdot P(\mathbf{X_A}|\alpha,\beta)\cdot P(\beta|\mathbf{X_A},\alpha)$. Each conditional probability on the right-hand side of the equation is estimated under the Maximal Likelihood Estimation (MLE), and thus we get the estimated probabilities of '$P(\alpha,\mathbf{X_A},\beta) = P(\alpha|\mathbf{X_A})\cdot P(\mathbf{X_A})\cdot P(\beta|\mathbf{X_A})$.'

Our probabilistic model is calculated under MLE. Lapata and Lascarides (2003) use this model as a psycholinguistic classification model for metonymic constructions. They find out that there is correlation between the human intuitive judgment and the measurement from the Lancaster-Oslo/Bergen corpus based on MLE.

2.3 Probabilistic Inference Model

We use the Bayes probability as a probabilistic inference computational model. This model has been widely accepted as WSD algorithm (Yarowsky 1992, Gale et al. 1992). We use the following equation in the Bayesian manner, where S indicates the set of senses; s denotes each of possible senses in S and W stands for a collocation for S; $S = \arg\max_{s \in S} \frac{p(W \mid s)P(s)}{P(W)}$

Lapata and Brew (2004) makes use of the naïve Bayes classifier for their research on the psycholinguistic model. Their research question is to find a systematic decision model of human intuition on verb class classification. We compare the decision behaviors projected from both human intuition and the three models considered in this paper.

3 Experiments

We gathered 7 homonyms from 1,000,000 word size Sejong corpus,[1] which is a Korean morphologically tagged corpus, also semantically annotated by experts. Homonyms we use are seven Korean words: *teulta*, *pae, sinpu, tari, keori, mak*, and *macta*. We first gathered sentences containing those homonyms and evaluated the data based on the definition in the standard Korean dictionary. The usage of each sense is varied and we chose the senses of over 20% in usage. Senses of each homonym are as follows; 'clear' and 'eat' for *teulta*, 'double' and 'abdomen' for *pae*, 'catholic father' and 'wedding bride' for *sinpu*, 'bridge' and 'leg' for *tari*, 'street' and 'distance' for *keori*, 'just' and 'often' for *mak*, and 'be hit' and 'harmonious' for *macta*.

[1] For details, see http://www.sejong.or.kr/english/index.html

Among 35,000 sentences we gathered, we used only 50 sentences for the experiments. The criterion for choosing them was to avoid the data propensity. We excluded a sentence containing a group of collocations with a notably high probability or frequency in the corpus. The probability of collocations that were selected lies between 0.0005 and 0.0000001.

The experiments were conducted on-line.[2] We controlled the experimental conditions; the place and the time. The subjects participated simultaneously in the experiment in a computer lab of a Korean university. Each subject was to respond within 20 second time limit to each item in the questionnaire. 30 undergraduate level students in all took part in the main experiment as volunteers unpaid.

We designed the questionnaires without the semantic priming effects (McNamara 2005). This experiment was not aimed to look into the semantic priming. We did a pretest in order to find any potential problems tied with the semantic priming, and then conducted the main test.

4 Discussion

Kappa statistics is used to check the inter-rater agreement between human annotators (Cohen 1960, Carletta 1995). The formula of Kappa statistics is $\kappa = \frac{P_a - P_e}{1 - P_e}$, where P_e measures the chance agreement and P_a represents for agreement rate between annotators. The Kappa value of our experimental results is 0.88 which proves the reliable agreements. In all of the items in the questionnaire, about 98% subjects agreed on one the given senses.

We initially measured the model's prediction on the basis of the 1,000,000 word size Sejong corpus. For instance, we gave the following context to the subjects as in (1), where *keori* is ambiguous among the senses of 'distance' and 'street':

(1) Keori-nun **eoceonhi** **hwalkie** **numcheo** iss-ess-ta.
Street-TOPIC still vitality flooded be-PAST-Ending
"The street was still flooded with the vitality."

The context in (1) contains the collocations of the content words in bold-face; *eoceonhi*(still), *hwalkie*(vitality), and *numceo*(flooding). The Boolean model searches each unique collocation group. We searched the corpus and found that *eoceonhi*(still), *hwalkie*(vitality) and *numceo*(flooding) are the unique entry of the collocations for the sense of 'street' and 'distance' has none. The numeric values of 'street' and 'distance' are 1.0 and 0.0, respectively. The probabilistic model predicts the occurrence estimations of each word. The Bayesian model measures the Bayesian predictions between two senses in this case.

The experiment on human intuition is calculated as the percentages against the whole group of subjects. If test result of (1) is such that the 'distance' group of subjects is none and the 'street' sense is 100%, the percentage of each sense is 0 and 100%.

We compare the percentage of the experimental result with the initial measurements predicted by each model in the corpus. The comparison between each of the measurements and the experiment is yielded by the Pearson correlation coefficient test. The results are shown in Table 1. The results show that the Boolean model shows the strongest correlation with the agreements results found in the human intuition experiments.

Table 1 Pearson correlation coefficient ($N = 100$)[3]

Model	Correlation	
	r	*p*
Boolean	0.686	0.001
MLE	0.674	0.001
Bayesian	0.517	0.001

[2] The website address is http://corpus.mireene.com/test.php.
[3] Statistics package used here is SPSS Korean Ver.12.0.

We also compared the precision of the sense classification made by each model with the sense classification of the experiments. The experimental result matched with the semantic annotation of the corpus. The Boolean model correctly predicted the decision at the rate of 95%, while the Bayesian classification was 90% correct. However, the MLE was only 75% correct.

It turned out that there was a case, an item in the questionnaire, where the Boolean model is not applicable. The context contains *keori* which is ambiguous among 'street' and 'distance' and contains collocations that occurs in both sense of *keori*. In this case, the Boolean model does not fit with the experimental results. The experimental result is 45:55 among two senses. The probabilistic model predicts 43:57, which is quite similar with the experimental result. This shows that the probabilistic model is more appropriate in this case.

5 Conclusion

In this paper, we compared three different WSD models with respect to the human intuition. Our experiment in this research is to look into the computational process of human intuition, and the results supports a strong correlation between the Boolean model and the human intuition.

Most WSD algorithms are based on the statistical behaviors of collocation. Their aim is to improve the precision and the recall, thus efficiently producing the best result of WSD. We observe in this research that the Boolean model best predicts the human behavior on semantic classification. Our way of looking into WSD is different, in reflecting a human behavior.

This experiment included the context where three models are applicable. But, we can think of a situation where some of three models are applicable or the hypo hypothetical situation where some models are tangibly interconnected. As a further study, we can extend the situation, looking into the human behavior.

References

1. Carletta, J. C.: Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics, Vol. 22 (1996) 249-254.
2. Cohen, J: A coefficient of agreement for nominal scales. Educational and Psychological Measurement, Vol. 20. (1960) 37–46.
3. Gale, W., K. Church, and D. Yarowsky: Using bilingual materials to develop word sense disambiguation methods: In the Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation (1992).
4. Harris, Z.: Distributional Structure. In Fodor Jerry and Katz Jerrold (ed.): The Structure of Language, Pritence-Hall (1964) 33-49.
5. Ide N. and J. Veronis: Introduction to the Special Issue on word Sense Disambiguation: The State of the Art, Computational Linguistics,Vol. 24. (1998) 1-40.
6. Lapata, M. and C. Brew: Verb Class Disambiguation Using Informative Priors. Computational Linguistics, Vol. 30, (2004) 45-73.
7. Lapata M. and A. Lascarides: A Probabilistic Account of Logical Metonymy, Computational Linguistics, Vol. 29, (2003) 44-88.
8. Mohammad S. and T. Pedersen: Combining Lexical and Syntactic Features for Supervised Word Sense Disambiguation: In the Proceedings of the Conference on Computational Natural Language Learning (CoNLL), , Boston, MA , May 6-7 (2004).
9. Yarowsky, D.: Word sense disambiguation using statistical models of Roget's categories trained on large corpora.: In the Proceedings of the 14th International Conference on Computational Linguistics, COLING'92, Nantes, France, August (1992).