# Potts Model on the Case Fillers for Word Sense Disambiguation

**Hiroya Takamura**
Tokyo Institute of Technology
takamura@pi.titech.ac.jp

**Manabu Okumura**
Tokyo Institute of Technology
oku@pi.titech.ac.jp

## Abstract

We propose a new method for word sense disambiguation for verbs. In our method, sense-dependent selectional preference of verbs is obtained through the probabilistic model on the lexical network. The mean-field approximation is employed to compute the state of the lexical network. The outcome of the computation is used as features for discriminative classifiers. The method is evaluated on the dataset of the Japanese word sense disambiguation.

## 1 Introduction

Polysemous words can be obstacles to many applications of natural language processing such as information retrieval, question answering, and machine translation. The task of distinguishing word senses given a token of a polysemous word and its context is often referred to as word sense disambiguation and has been studied by many researchers (Agirre and Edmonds, 2006). Among a number of word sense disambiguation tasks including noun sense disambiguation, adjective sense disambiguation, and named entity disambiguation, we focus on the verb sense disambiguation with the supervised setting, where we are supposed to construct a classifier given labeled training instances.

It is learned in the previous work that case fillers are often good clues for the verb sense disambiguation. Consider, for example, the following two sentences:

1. "He drove a car to the next town."
2. "He drove the dogs away."

The "drive" in Sentence 1 means "operate (a vehicle)", while "drive" in Sentence 2 means "urge (something to move)". Although there might be long contexts to these instances, looking at the case fillers "car" and "dog" alone would lead to the correct interpretations of the meanings. This kind of preference on nouns as case fillers is called $selectional\ preference$, which in this case depends on sense. It is sometimes impractical, however, to expect that the training dataset covers all the nouns as case fillers, if case fillers of the target verb are diverse. The purpose of this article is to propose a method for propagating the information on the case fillers in the training dataset to other nouns, so that the sense of polysemous words is correctly disambiguated. In our method, the information propagation is implemented as estimation of the state of the probabilisitc model on the lexical network. One advantage of our method is that we can overcome the difficulty caused by the noise contained in the lexical network.

## 2 Related Work

Recent work on general word sense disambiguation is summarized in the book edited by Agirre and Edmonds (2006). We focus on the work of the verb sense disambiguation.

The idea of using case frames for the verb sense disambiguation is not novel, and dates back to 90's. Fujii et al. (1998) proposed a method for the verb sense disambiguation, which is based on the $k$-nearest neighbors. In their method, each instance is represented as a case frame and the similarity of two instances is calculated as a weighted sum of the similarities of case fillers. Fujii et al. also proposed a framework of active learning.

To disambiguate verb senses, Chen and Palmer (2009) proposed to use linguistic and semantic features including the voice of the given sentence, the presence of a PP adjunct, and the named entity tags. Dligach and Palmer (2008) proposed to use co-occurrence with other verbs as features. Wagner et al. (2009) proposed to use verb clusters generated with statistics on verb subcategorization.

*Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 1382–1386,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

## 3 Potts Model

We introduce the probability model that we will use for our task. This model gives the probability distribution over a set of nodes associated with random variables, where some pairs of variables are dependent on each other. If the variables can have more than two values and there is no order relation between the values, the network comprised of such variables is called *Potts* model (Wu, 1982), which has been used in applications such as image restoration (Tanaka and Morita, 1996) and rumor transmission (Liu et al., 2001).

Suppose a network consisting of nodes and weighted edges is given. Let $c$ denote the value of a node, and $w_{ij}$ the weight between $i$ and $j$. Energy function $H(\mathbf{c})$ is represented as

$$H(\mathbf{c}) = -\beta \sum_{ij} w_{ij}\delta(c_i, c_j) - \alpha \sum_{i \in L} \delta(c_i, a_i), \quad (1)$$

where $\beta$ is a constant called *the inverse-temperature*, $L$ is the index set for the observed variables, $a_i$ is the value of an observed variable $i$, and $\alpha$ is a positive constant representing a weight on labeled data. $\delta$ returns 1 if two arguments are equal to each other, 0 otherwise. The state is penalized if $c_i$ $(i \in L)$ is different from $a_i$. The probability distribution of the network is represented as $P(\mathbf{c}) = \exp\{-H(\mathbf{c})\}/Z$, where $Z$ is a normalization factor.

Instead of minimizing $H(\mathbf{c})$, we attempt to minimize the free energy, which is defined to be the sum of $H(\mathbf{c})$ and the negative entropy. However, this minimization is computationally hard. We hence resort to the mean-field approximation method (Nishimori, 2001), in which $P(\mathbf{c})$ is replaced by factorized function $\rho(\mathbf{c}) = \prod_i \rho_i(c_i)$. The evergy function with the factorized probability function is called *variational free energy*:

$$\begin{aligned} F(\mathbf{c}) &= \sum_{\mathbf{c}} \rho(\mathbf{c}) H(\mathbf{c}) - \sum_{\mathbf{c}} -\rho(\mathbf{c}) \log \rho(\mathbf{c}) \\ &= -\alpha \sum_i \sum_{c_i} \rho_i(c_i)\delta(c_i, a_i) \\ &\quad -\beta \sum_{ij} \sum_{c_i, c_j} \rho_i(c_i)\rho_j(c_j) w_{ij}\delta(c_i, c_j) \\ &\quad - \sum_i \sum_{c_i} -\rho_i(c_i) \log \rho_i(c_i). \quad (2) \end{aligned}$$

By minimizing $F(\mathbf{c})$ under the condition that $\forall i, \sum_{c_i} \rho_i(c_i) = 1$, we obtain the following fixed point equation for $i \in L$:

$$\rho_i(c) = \frac{\exp(\alpha\delta(c, a_i) + \beta \sum_j w_{ij}\rho_j(c))}{\sum_n \exp(\alpha\delta(n, a_i) + \beta \sum_j w_{ij}\rho_j(n))}. \quad (3)$$

The fixed point equation for $i \notin L$ can be obtained by removing $\alpha\delta(c, a_i)$ from above. This fixed point equation is solved by an iterative computation. In the actual implementation, we represent $\rho_i$ with a linear combination of the discrete Tchebycheff polynomials (Tanaka and Morita, 1996). Details on the Potts model and its computation can be found in the literature (Nishimori, 2001).

## 4 Proposed Method

### 4.1 Construction of Lexical Networks

We follow the work by Takamura et al. (2005) to construct a lexical network. We link two words if one word appears in the gloss of the other. Each link belongs to one of two groups: the same-orientation links $SL$ and the different-orientation links $DL$. If a negation word appears in the gloss of an entry word, the words after the negation word are linked to the entry word with $DL$. Otherwise, those words are linked with $SL$. In case of Japanese, the auxiliaries "nai" and "nu" are regarded as negation words.

We next set weights $W = (w_{ij})$ to links :

$$w_{ij} = \begin{cases} \frac{1}{\sqrt{d(i)d(j)}} & (l_{ij} \in SL) \\ -\frac{1}{\sqrt{d(i)d(j)}} & (l_{ij} \in DL) \\ 0 & otherwise \end{cases}, \quad (4)$$

where $l_{ij}$ denotes the link between word $i$ and word $j$, and $d(i)$ denotes the degree of word $i$, which indicates the number of words linked with word $i$. We call this network *the gloss network (G)*. We construct another network, *the thesaurus network (T)*, by linking synonyms and hypernyms in a thesaurus. We also merge the two networks above to construct another network *the gloss-thesaurus network (GT)*.

### 4.2 Use of Potts Model

We estimate the tendency of each word to be the case filler for a verb sense. For example, "car" would have a high tendency to be the case filler for "drive" with the sense "operate (a vehicle)". Such tendency is measured as the probability $P_i(c)$ over senses $c$ for noun $n_i$. We will use the Potts model for this purpose. Namely, the local approximate probability $\rho_i(c)$, equivalently the mariginal probability $\sum_{\boldsymbol{c}} \rho(\boldsymbol{c})$, is regarded as probability $P_i(c)$.

The values of each random variable are senses of the target verb. For each case and each verb, we estimate the probability of the sense assignments on the whole set of nodes by means of the mean-field approximation. The case fillers of the training instances are used as observed variables, with their index set being $L$ in Equation (2). $P_i(c)$ is estimated for each case.

In our experiments on Japanese, surface cases are employed: $wo$ (accusative), $ga$ (nominative), $ni$ (dative/locative), $de$ (locative/instrumental), $no$ (genitive/others), $e$ (locative/illative), $to$ (comitative), $kara$ (elative), $yori$ (comparative), $made$ (terminative). Although our model is also applicable to deep cases, we focus on surface cases since deep case recognition itself is a challenging task.

### 4.3 Estimation of $\beta$

In some pieces of previous work (Takamura et al., 2005; Takamura et al., 2007), it has been shown that the optimal $\beta$ can be obtained by estimating the critical temperature, at which phase transition occurs from *paramagnetic phase* (variables are randomly oriented) to *ferromagnetic phase* (most of the variables have the same value). We follow these pieces of previous work.

In practice, when the maximum of the spatial averages of the approximated probabilities $\max_c \sum_i \rho_i(c)/N$ exceeds a threshold during increasing $\beta$, we consider that the phase transition has occurred. We select the value of $\beta$ slightly before the phase transition.

### 4.4 Discriminative Training

Probability $\rho_i(c)$ is expected to be a strong clue for sense disambiguation. However, it is not clear how we can effectively use the probabilities of different cases $c$; $\rho_i(c)$ for some cases would be reliable, while some others less reliable. In addition, the word tokens that appear before and after the target word also give evidence for senses. We therefore use a discriminative approach to construct a classifier with those various clues as features. Support vector machines are employed in this work, although other classifiers are also applicable.

Note that even if a case filler in the test instance did not appear in the training data, the feature $\rho_i(c)$ corresponding to this case filler conveys the information on the selectional preference of the verb against the case filler.

## 5 Experiments

### 5.1 Experimental Settings

The proposed method for verb sense disambiguation is evaluated on the white paper part of BCCWJ corpus, the first balanced corpus of contemporary written Japanese (Maekawa, 2008), which was also used as a test set for SemEval-2 Japanese word sense disambiguation task (Okumura et al., 2010). The dataset used in this research was created by the preliminary annotation for SemEval-2. The senses are defined in the Iwanami Kokugo Jiten (Nishio et al., 1994), a Japanese dictionary. Among three levels of sense IDs defined in this dictionary, the middle-level sense was used in the empirical evaluation, which is the same level of senses used in the SemEval-2. From the dataset above, we selected most ambiguous 14 verbs whose empirical sense distributions (i.e., estimated with the maximum likelihood principle) have a high entropy and appear more than 100 times in the dataset. The statistics of this dataset is shown in the middle columns of Table 1. 5-fold cross-validation was employed for each verb.

TinySVM version 0.09[1] was used for SVM training and classification. The linear kernel was used as a kernel function of SVM. The soft-margin parameter $C$ indicating the tradeoff between the model complexity and the training error was tuned to the best value among $0.01, 0.1, 1, 10$ for each method. The glosses of the Iwanami Kokugo Jiten Japanese dictionary are used to construct a lexical network $G$ (gloss network). Japanese WordNet version 0.9 (Bond et al., 2009) was used to construct a lexical network $T$ (thesaurus network). The numbers of nodes in $G$, $T$, and $GT$ are respectively 35225, 66218, and 87038. Due to a large number of polysemous words in the thesaurus, many synsets of Japanese WordNet are connected.

Two baselines are evaluated on the test datasets. Baseline 1 is SVMs trained with basic features: the target verb itself and 3 words before and after the target verb. Baseline 2 is also SVMs trained with the basic features above and the case filler features: the nouns that appear as case fillers.

The proposed methods are SVMs trained with the basic features, the case filler features, and the lexical network features: the probability $\rho_i(c)$ introduced in Section 4 for each case and each sense $c$ when $w_i$ appears as a case filler.

[1] http://chasen.org/~taku/software/TinySVM/

Table 1: Classification accuracy on $hard$ verbs (%)

| target word | English translation | # of instances | # of senses | baselines 1 | baselines 2 | $G$ | $T$ | $GT$ |
|---|---|---|---|---|---|---|---|---|
| $ataru$ | hit, correspond | 463 | 7 | 83.2 | 85.3 | 87.9 | 88.1 | 87.7 |
| $dasu$ | take out, cause | 173 | 3 | 73.4 | 75.1 | 74.6 | 79.2 | 75.1 |
| $deru$ | go out, appear | 189 | 3 | 86.2 | 86.8 | 89.4 | 87.8 | 87.8 |
| $kiku$ | listen, be effective | 141 | 2 | 87.2 | 87.2 | 87.2 | 86.5 | 87.2 |
| $susumu$ | proceed, advance | 931 | 2 | 81.8 | 83.1 | 84.2 | 83.5 | 84.4 |
| average | – | 379.4 | 3.4 | 82.0 | 83.5 | 84.7 | 84.6 | 84.8 |

Table 2: Average classification accuracy on $easy$ verbs (%)

|  | # of instances | # of senses | baselines 1 | baselines 2 | $G$ | $T$ | $GT$ |
|---|---|---|---|---|---|---|---|
| average on $easy$ verbs | 1216.9 | 3.0 | 94.6 | 94.8 | 95.0 | 95.1 | 95.0 |
| total average | 917.8 | 3.1 | 93.1 | 93.4 | 93.7 | 93.8 | 93.7 |

The classification accuracy, i.e., the number of correctly classified instances divided by the total number of instances, is employed as evaluation measure. Averages in this paper are micro averages. The methods are evaluated first for $hard$ verbs, for which the classification accuracy of the baseline method is less than 90 %. There are 5 $hard$ verbs; the other 9 verbs are $easy$ verbs, for which the classification accuracy is already equal to or better than 90 %. Note that the proposed method aims at $hard$ verbs as mentioned in Section 1, although we would like the proposed method not to degrade the classification performance for $easy$ verbs.

### 5.2 Results

The classification result for $hard$ verbs is shown in Table 1. Baseline 2 is better than baseline 1, meaning that the simple information on the word as a case filler improves the classification performance. The proposed methods on lexical networks $G$, $T$, and $GT$ mostly outperformed baselines with a few exceptions. On average, the proposed method on $GT$ increases the classification accuracy by 2.8 points compared with the baseline 1, 1.3 points compared with the baseline 2. This result shows that the information propagation on the lexical network offers useful clues for verb sense disambiguation. Table 2 shows that the proposed method is at least comparable to the baselines when it is applied to $easy$ verbs. The accuracy for each $easy$ verb was omitted from the table due to the space limitation, and only the average values are written in the table. These results also show that the difference of lexical networks does not have a significant effect on the average classification performance.

In order to gain more intuitive understanding on the method, we give an example of the computational result for case $wo$ (accusative) of verb $dasu$. We used all of the 173 instances that are available. Nouns that have high probability $\rho_i(c)$ for $c = sense1$ (take out, let out, send, pay) are, for example, price, application, permission, demand, request, wish, command, permit, and certificate. Nouns that have high $\rho_i(c)$ for $c = sense2$ (show) are, for example, speed, advance, agility, effect, driving force, breakthrough, and activity.

## 6 Conclusion

We proposed a new method for word sense disambiguation for verbs. In our method, sense-dependent selectional preference of verbs was obtained through the probabilistic model on the lexical network. The mean-field approximation was employed to compute the state of the lexical network. The outcome of the computation was used as features for discriminative classifiers. The method is evaluated on the dataset of the Japanese word sense disambiguation.

Future work includes the use of words that are not on the lexical network, the incorporation of interactions between multiple cases, and theoretical study of the Potts model for lexical network.

## References

Eneko Agirre and Philip Glenny Edmonds, editors. 2006. *Word Sense Disambiguation: Algorithms And Applications*. Springer-Verlag.

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the japanese wordnet. In *Proceedings of the 7th Workshop on Asian Language Resources (in conjunction with ACL-IJCNLP 2009)*.

Jinying Chen and Martha S. Palmer. 2009. Improving English verb sense disambiguation performance with linguistically motivated features and clear sense distinction boundaries. *Language Resources and Evaluation*, 43:181–208.

Dmitriy Dligach and Martha Palmer. 2008. Novel semantic features for verb sense disambiguation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pages 29–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. 1998. Selective sampling for example-based word sense disambiguation. *Computational Linguistics*, 24(4):573–597.

Zhongzhu Liu, Jun Luo, and Chenggang Shao. 2001. Potts model for exaggeration of a simple rumor transmitted by recreant rumormongers. *Physical Review E*, 64:046134,1–046134,9.

Kikuo Maekawa. 2008. Balanced corpus of contemporary written japanese. In *Proceedings of the 6th Workshop on Asian Language Resources*, pages 101–102.

Hidetoshi Nishimori. 2001. *Statistical Physics of Spin Glasses and Information Processing*. Oxford University Press.

Minoru Nishio, Etsutaro Iwabuchi, and Shizuo Mizutani. 1994. *Iwanami Japanese Dictionary (5th edition)*. Iwanami-shoten.

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. 2010. Semeval-2010 task: Japanese wsd. In *Proceedings of the 5th International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 69–74.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 133–140.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2007. Extracting semantic orientations of phrases from dictionary. In *Proceedings of the Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT2007)*, pages 292–299.

Kazuyuki Tanaka and Tohru Morita. 1996. Application of cluster variation method to image restoration problem. In J.L. Morán-López and J.M. Sanchez, editors, *Theory and Applications of the Cluster Variation and Path Probability Methods*, pages 353–373. Plenum Press, New York.

Wiebke Wagner, Helmut Schmid, and Sabine Schulte im Walde. 2009. Verb sense disambiguation using a predicate-argument-clustering model. In *Proceedings of the CogSci Workshop on Distributional Semantics beyond Concrete Concepts*.

Fa-Yueh Wu. 1982. The potts model. *Reviews of Modern Physics*, 54(1):235–268.