Density Maximization in Context-Sense Metric Space for All-words WSD

Koichi Tanigaki[†‡] **Mitsuteru Shiba**[†] **Tatsuji Munaka**[†] **Yoshinori Sagisaka**[‡]
† Information Technology R&D Center, Mitsubishi Electric Corporation
5-1-1 Ofuna, Kamakura, Kanagawa 247-8501, Japan
‡ Global Information and Telecommunication Institute, Waseda University
1-3-10 Nishi-Waseda, Shinjuku-ku, Tokyo 169-0051, Japan

Abstract

This paper proposes a novel *smoothing model* with a *combinatorial optimization scheme* for all-words word sense disambiguation from untagged corpora. By generalizing discrete senses to a continuum, we introduce a smoothing in context-sense space to cope with *data-sparsity* resulting from a large variety of linguistic context and sense, as well as to exploit *sense-interdependency* among the words in the same text string. Through the smoothing, all the optimal senses are obtained at one time under maximum marginal likelihood criterion, by competitive probabilistic kernels made to reinforce one another among nearby words, and to suppress conflicting sense hypotheses within the same word. Experimental results confirmed the superiority of the proposed method over conventional ones by showing the better performances beyond most-frequent-sense baseline performance where none of SemEval-2 unsupervised systems reached.

1 Introduction

Word Sense Disambiguation (WSD) is a task to identify the intended sense of a word based on its context. *All-words WSD* is its variant, where all the unrestricted running words in text are expected to be disambiguated. In the all-words task, all the senses in a dictionary are potentially the target destination of classification, and purely supervised approaches inherently suffer from *data-sparsity* problem. The all-words task is also characterized by *sense-interdependency* of target words. As the target words are typically taken from the same text string, they are naturally expected to be interrelated. Disambiguation of a word should affect other words as an important clue.

From such characteristics of the task, knowledge-based unsupervised approaches have been extensively studied. They compute dictionary-based sense similarity to find the most related senses among the words within a certain range of text. (For reviews, see (Agirre and Edmonds, 2006; Navigli, 2009).) In recent years, graph-based methods have attracted considerable attentions (Mihalcea, 2005; Navigli and Lapata, 2007; Agirre and Soroa, 2009). On the graph structure of lexical knowledge base (LKB), random-walk or other well-known graph-based techniques have been applied to find mutually related senses among target words. Unlike earlier studies disambiguating word-by-word, the graph-based methods obtain *sense-interdependent* solution for target words. However, those methods mainly focus on modeling sense distribution and have less attention to contextual smoothing/generalization beyond immediate context.

There exist several studies that enrich immediate context with large corpus statistics. McCarthy et al. (2004) proposed a method to combine sense similarity with distributional similarity and configured predominant sense score. Distributional similarity was used to weight the influence of context words, based on large-scale statistics. The method achieved successful WSD accuracy. Agirre et al. (2009) used a k-nearest words on distributional similarity as context words. They apply a LKB graph-based WSD to a target word together with the distributional context words, and showed that it yields better results on a domain dataset than just using immediate context words. Though these

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 884–893,
Sofia, Bulgaria, August 4-9 2013.

studies are word-by-word WSD for target words, they demonstrated the effectiveness to enrich immediate context by corpus statistics.

This paper proposes a *smoothing model* that integrates dictionary-based semantic similarity and corpus-based context statistics, where a *combinatorial optimization scheme* is employed to deal with sense interdependency of the all-words WSD task. The rest of this paper is structured as follows. We first describe our smoothing model in the following section. The combinatorial optimization method with the model is described in Section 3. Section 4 describes a specific implementation for evaluation. The evaluation is performed with the SemEval-2 English all-words dataset. We present the performance in Section 5. In Section 6 we discuss whether the intended context-to-sense mapping and the sense-interdependency are properly modeled. Finally we review related studies in Section 7 and conclude in Section 8.

2 Smoothing Model

Let us introduce in this section the basic idea for modeling context-to-sense mapping. The distance (or similarity) metrics are assumed to be given for context and for sense. A specific implementation of these metrics is described later in this paper, for now the context metric is generalized with a distance function $d_{\mathrm{x}}(\cdot,\cdot)$ and the sense metric with $d_{\mathrm{s}}(\cdot,\cdot)$. Actually these functions may be arbitrary ones that accept two elements and return a positive real number.

Now suppose we are given a dataset concerning N number of target words. This dataset is denoted by $X = \{x_i\}_{i=1}^{N}$, where x_i corresponds to the context of the i-th word but not the word by itself. For each x_i, the intended sense of the word is to be found in a set of sense candidates $S_i = \{s_{ij}\}_{j=1}^{M_i} \subseteq S$, where M_i is the number of sense candidates for the i-th word, S is the whole set of sense inventories in a dictionary. Let the two-tuple $h_{ij} = (x_i, s_{ij})$ be the hypothesis that the intended sense in x_i is s_{ij}. The hypothesis is an element of the direct product $H = X \times S$. As (X, d_{x}) and (S, d_{s}) each composes a metric space, H is also a metric space, provided a proper distance definition with d_{x} and d_{s}.

Here, we treat the space H as a continuous one, which means that we assume *the relationship between context and sense can be generalized in continuous fashion*. In natural language processing, continuity has been sometimes assumed for linguistic phenomena including word context for corpus based WSD. As for classes or senses, it may not be a common assumption. However, when the classes for all-words WSD are enormous, fine-grained, and can be associated with distance, we can rather naturally assume the continuity also for senses. According to the nature of continuity, once given a hypothesis h_{ij} for a certain word, we can extrapolate the hypothesis for another word of another sense $h_{i'j'} = (x_{i'}, s_{i'j'})$ sufficiently close to h_{ij}. Using a Gaussian kernel (Parzen, 1962) as a smoothing model, the probability density extrapolated at $h_{i'j'}$ given h_{ij} is defined by their distance as follows:

$$\mathcal{K}(h_{ij}, h_{i'j'}) \equiv \frac{1}{2\pi\sigma_{\mathrm{x}}\sigma_{\mathrm{s}}} \exp\left[-\frac{{d_{\mathrm{x}}}^2(x_i, x_{i'})}{2{\sigma_{\mathrm{x}}}^2} - \frac{{d_{\mathrm{s}}}^2(s_{ij}, s_{i'j'})}{2{\sigma_{\mathrm{s}}}^2} \right], \tag{1}$$

where σ_{x} and σ_{s} are parameters of positive real number $\sigma_{\mathrm{x}}, \sigma_{\mathrm{s}} \in \mathbb{R}^+$ called *kernel bandwidths*. They control the smoothing intensity in context and in sense, respectively.

Our objective is to determine the optimal sense for all the target words simultaneously. It is essentially a 0-1 integer programing problem, and is not computationally tractable. We relax the integer constraints by introducing a *sense probability* parameter π_{ij} corresponding to each h_{ij}. π_{ij} denotes the probability by which h_{ij} is true. As π_{ij} is a probability, it satisfies the constraints $\forall i\ \sum_j \pi_{ij} = 1$ and $\forall i, j\ 0 \leq \pi_{ij} \leq 1$. The probability density extrapolated at $h_{i'j'}$ by a *probabilistic* hypothesis h_{ij} is given as follows:

$$\mathcal{Q}_{ij}(h_{i'j'}) \ \propto\ \pi_{ij}\,\mathcal{K}(h_{ij}, h_{i'j'}). \tag{2}$$

The proposed model is illustrated in Figure 1. Due to the limitation of drawing, both the context metric space and the sense metric space are drawn schematically as 1-dimensional spaces (axes), actually arbitrary metric spaces similarity-based or feature-based are applicable. The product metric space of the context metric space and the sense metric space composes a hypothesis space. In the hypothesis space, n sense hypotheses for a certain word is represented as n points on the hyperplane that spreads across the sense metric space. The two small circles in the middle of the figure represent the two sense hypotheses for a single word. The position of a hypothesis represents which sense is assigned to the current word in

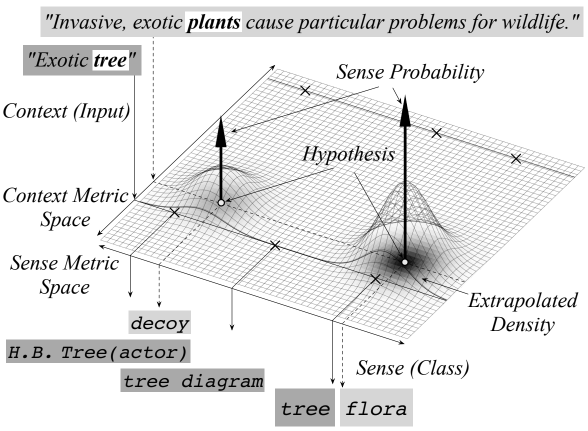


Figure 1: Proposed probability distribution model for context-to-sense mapping space.

what context. The upward arrow on a hypothesis represents the magnitude of its probability.

Centered on each hypotheses, a Gaussian kernel is placed as a smoothing model. It extrapolates the hypotheses of other words around it. In accordance with geometric intuition, intensity of extrapolation is affected by the distance from a hypothesis, and by the probability of the hypothesis by itself. Extrapolated probability density is represented by shadow thickness and surface height. If there is another word in nearby context, the kernels can validate the sense of that word. In the figure, there are two kernels in the context *"Invasive, exotic ..."*. They are two competing hypothesis for the senses `decoy` and `flora` of the word *plants*. These kernels affect the senses of another ambiguous word *tree* in nearby context *"Exotic ..."*, and extrapolate the most at the sense `tree` nearby `flora`. The extrapolation has non-linear effect. It affects little to the word far away in context or in sense as is the case for the background word in the figure. Strength of smoothing is determined by kernel bandwidths. Wider bandwidths bring stronger effect of generalization to further hypotheses, but too wide bandwidths smooth out detailed structure. The bandwidths are the key for disambiguation, therefore they are to be optimized on a dataset together with sense probabilities.

3 Simultaneous Optimization of All-words WSD

Given the smoothing model to extrapolate the senses of other words, we now make its instances interact to obtain the optimal combination of senses for all the words.

3.1 Likelihood Definition

Let us first define the likelihood of model parameters for a given dataset. The parameters consist of a context bandwidth σ_{x}, a sense bandwidth σ_{s}, and sense probabilities π_{ij} for all i and j. For convenience of description, the sense probabilities are all together denoted as a vector $\boldsymbol{\pi} = (\ldots, \pi_{ij}, \ldots)^\top$, in which actual order is not the matter.

Now remind that our dataset $X = \{x_i\}_{i=1}^N$ is composed of N instances of *unlabeled* word context. We consider all the mappings from context to sense are latent, and find the optimal parameters by maximizing marginal pseudo likelihood based on probability density. The likelihood is defined as follows:

$$\mathcal{L}(\boldsymbol{\pi}, \sigma_{\mathrm{x}}, \sigma_{\mathrm{s}}; X) \equiv \ln \prod_i \sum_j \pi_{ij} \mathcal{Q}(h_{ij}), \quad (3)$$

where $\prod_i$ denotes the product over $x_i \in X$, $\sum_j$ denotes the summation over all possible senses $s_{ij} \in S_i$ for the current i-th context. $\mathcal{Q}(h_{ij})$ denotes the probability density at h_{ij}. We compute $\mathcal{Q}(h_{ij})$ using *leave-one-out cross-validation* (LOOCV), so as to prevent kernels from overfitting to themselves, as follows:

$$\mathcal{Q}(h_{ij}) \quad (4)$$
$$\equiv \frac{1}{N - N_i} \sum_{i' :\, w_{i'} \neq w_i} \sum_{j'} \pi_{i'j'} \mathcal{K}(h_{ij}, h_{i'j'}),$$

where N_i denotes the number of occurrences of a word type w_i in X, and $\sum_{i' :\, w_{i'} \neq w_i}$ denotes the summation over $x_{i'} \in X$ except the case that the word type $w_{i'}$ equals to w_i. $\sum_{j'}$ denotes the summation over $s_{i'j'} \in S_{i'}$. We take as the unit of LOOCV not a word instance but a word type, because the instances of the same word type invariably have the same sense candidates, which still cause over-fitting when optimizing the sense bandwidth.

3.2 Parameter Optimization

We are now ready to calculate the optimal senses. The optimal parameters $\boldsymbol{\pi}^*, \sigma_{\mathrm{x}}^*, \sigma_{\mathrm{s}}^*$ are obtained by maximizing the likelihood $\mathcal{L}$ subject to the constraints on $\boldsymbol{\pi}$, that is $\forall i \ \sum_j \pi_{ij} = 1$ [1]. Using the Lagrange multipliers $\{\lambda_i\}_{i=1}^N$ for every i-th constraint, the solution for the constrained maximiza-

[1]It is guaranteed that the other constraints $\forall i, j \; 0 \leq \pi_{ij} \leq 1$ are satisfied according to Equation (7).

tion of $\mathcal{L}$ is obtained as the solution for the equivalent unconstrained maximization of $\check{\mathcal{L}}$ as follows:

$$\boldsymbol{\pi}^*, \sigma_\mathrm{x}^*, \sigma_\mathrm{s}^* \;=\; \operatorname*{arg\,max}_{\boldsymbol{\pi},\, \sigma_\mathrm{x},\, \sigma_\mathrm{s}} \check{\mathcal{L}}, \tag{5}$$

where

$$\check{\mathcal{L}} \;\equiv\; \mathcal{L} + \sum_i \lambda_i \left(\sum_j \pi_{ij} - 1 \right). \tag{6}$$

When we optimize the parameters, the first term of Equation (6) in the right-hand side acts *to reinforce* nearby hypotheses among different words, whereas the second term acts *to suppress* conflicting hypotheses of the same word.

Taking $\nabla \check{\mathcal{L}} = 0$, erasing λ_i, and rearranging, we obtain the optimal parameters as follows:

$$\pi_{ij} \;=\; \frac{\sum_{\substack{i',j' \\ w_{i'} \neq w_i}} \mathcal{R}^{ij}_{i'j'} + \sum_{\substack{i',j' \\ w_{i'} \neq w_i}} \mathcal{R}^{i'j'}_{ij}}{1 \;+\; \sum_j \sum_{\substack{i',j' \\ w_{i'} \neq w_i}} \mathcal{R}^{i'j'}_{ij}} \tag{7}$$

$${\sigma_\mathrm{x}}^2 \;=\; \frac{1}{N} \sum_{\substack{i,i',j,j' \\ w_{i'} \neq w_i}} \mathcal{R}^{ij}_{i'j'}\, {d_\mathrm{x}}^2(x_i, x_{i'}) \tag{8}$$

$${\sigma_\mathrm{s}}^2 \;=\; \frac{1}{N} \sum_{\substack{i,i',j,j' \\ w_{i'} \neq w_i}} \mathcal{R}^{ij}_{i'j'}\, {d_\mathrm{s}}^2(s_{ij}, s_{i'j'}), \tag{9}$$

where $\mathcal{R}^{ij}_{i'j'}$ denotes the *responsibility* of $h_{i'j'}$ to h_{ij}: the ratio of total expected density at h_{ij}, taken up by the expected density extrapolated by $h_{i'j'}$, normalized to the total for x_i be 1. It is defined as

$$\mathcal{R}^{ij}_{i'j'} \;\equiv\; \frac{\pi_{ij}\, \mathcal{Q}_{i'j'}(h_{ij})}{\sum_j \pi_{ij}\, \mathcal{Q}(h_{ij})}. \tag{10}$$

$\mathcal{Q}_{i'j'}(h_{ij})$ denotes the probability density at h_{ij} extrapolated by $h_{i'j'}$ alone, defined as follows:

$$\mathcal{Q}_{i'j'}(h_{ij}) \;\equiv\; \frac{1}{N - N_i}\, \pi_{i'j'} \mathcal{K}(h_{ij}, h_{i'j'}). \tag{11}$$

Intuitively, Equations (7)-(9) are interpreted as follows. As for Equation (7), the right-hand side of the equation can be divided as the left term and the right term both in the numerator and in the denominator. The left term requires π_{ij} to agree with the ratio of responsibility *of the whole to* h_{ij}. The right term requires π_{ij} to agree with the ratio of responsibility *of* h_{ij} *to the whole*. As for Equation (8), (9), the optimal solution is the mean squared distance in context, and in sense, weighted by responsibility.

To obtain the actual values of the optimal parameters, EM algorithm (Dempster et al., 1977) is applied. This is because Equations (7)-(9) are circular definitions, which include the objective parameters implicitly in the right hand side, thus the solution is not obtained analytically. EM algorithm is an iterative method for finding maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables. Applying the EM algorithm to our model, we obtain the following steps:

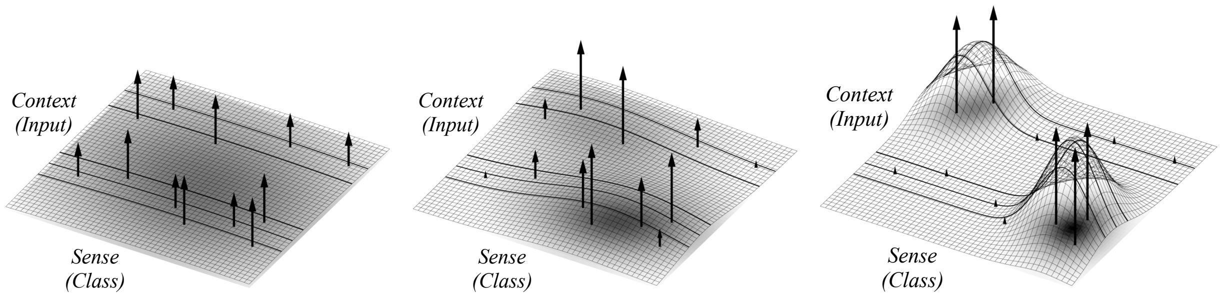
Step 1. Initialization: Set initial values to $\boldsymbol{\pi}$, σ_x, and σ_s. As for sense probabilities, we set the uniform probability in accordance with the number of sense candidates, thereby $\pi_{ij} \leftarrow |S_i|^{-1}$, where $|S_i|$ denotes the size of S_i. As for bandwidths, we set the mean squared distance in each metric; thereby ${\sigma_\mathrm{x}}^2 \leftarrow N^{-1} \sum_{i,i'} {d_\mathrm{x}}^2(x_i, x_{i'})$ for context bandwidth, and ${\sigma_\mathrm{s}}^2 \leftarrow (\sum_i |S_i|)^{-1} \sum_{i,i'} \sum_{j,j'} {d_\mathrm{s}}^2(s_{ij}, s_{i'j'})$ for sense bandwidth.

Step 2. Expectation: Using the current parameters $\boldsymbol{\pi}$, σ_x, and σ_s, calculate the responsibilities $\mathcal{R}^{ij}_{i'j'}$ according to Equation (10).

Step 3. Maximization: Using the current responsibility $\mathcal{R}^{ij}_{i'j'}$, update the parameters $\boldsymbol{\pi}$, σ_x, and σ_s, according to Equation (7)-(9).

Step 4. Convergence test: Compute the likelihood. If its ratio to the previous iteration is sufficiently small, or predetermined number of iterations has been reached, then terminate the iteration. Otherwise go back to Step 2.

To visualize how it works, we applied the above EM algorithm to pseudo 2-dimensional data. The results are shown in Figure 2. It simulates WSD for an $N = 5$ words dataset, whose contexts are depicted by five lines. The sense hypotheses are depicted by twelve upward arrows. At the base of each arrow, there is a Gaussian kernel. Shadow thickness and surface height represents the composite probability distribution of all the twelve kernels. Through the iterative parameter update, sense probabilities and kernel bandwidths were optimized to the dataset. Figure 2(a) illustrates the initial status, where all the sense hypothesis are equivalently probable, thus they are in the most ambiguous status. Initial bandwidths are set to the mean squared distance of all the hypotheses pairs,



(a) Initial status. (b) Status after the 7th iteration. (c) Converged status after 25 iterations.

Figure 2: Pseudo 2D data simulation to visualize the dynamics of the proposed simultaneous all-words WSD with ambiguous five words and twelve sense hypotheses. (There are twelve Gaussian kernels at the base of each arrow, though the figure shows just their composite distribution. Those kernels reinforce and compete one another while being fitted their affecting range, and finally settle down to the most consistent interpretation for the words with appropriate generalization. For the dynamics with an actual dataset, see Figure 5.)

which is rather broad and makes kernels strongly smoothed, thus the model captures general structure of space. Figure 2(b) shows the status after the 7th iteration. Bandwidths are shrinking especially in context, and two context clusters, so to speak, two usages, are found. Figure 2(c) shows the status of convergence after 25 iterations. All the arrow lengths incline to either 1 or 0 along with their neighbors, thus all the five words are now disambiguated.

Note that this is not the conventional clustering of observed data. If, for instance, the Gaussian mixture clustering of 2-mixtures is applied to the positions of these hypotheses, it will find the clusters just like Figure 2(b) and will stop. The cluster centers are located at the means of hypotheses including miscellaneous alternatives not intended, thus the estimated probability distribution is, roughly speaking, offset toward the center of WordNet, which is not what we want. In contrast, the proposed method proceeds to Figure 2(c) and finds clusters *in the data after conflicting data is erased*. This is because our method is aiming at modeling not the disambiguation of cluster-memberships but the disambiguation of senses for each word.

4 Metric Space Implementation

So far, we have dealt with general metrics for context and for sense. This section describes a specific implementation of those metrics employed in the evaluation. We followed the previous study by McCarthy et al. (2004), (2007), and implemented a type-based WSD. The context of word instances are tied to the distributional context of the word type in a large corpus. To calculate sense similarities, we used the WordNet similarity package by Pedersen et al. (2004), version 2.05. Two measures proposed by Jiang and Conrath (1997) and Lesk (1986) were examined, which performed best in the previous study (McCarthy et al., 2004).

Distributional similarity (Lin, 1998) was computed among target words, based on the statistics of the test set and the background text provided as the official dataset of the SemEval-2 English all-words task (Agirre et al., 2010). Those texts were parsed using RASP parser (Briscoe et al., 2006) version 3.1, to obtain grammatical relations for the distributional similarity, as well as to obtain lemmata and part-of-speech (POS) tags which are required to look up the sense inventory of WordNet. Based on the distributional similarity, we just used k-nearest neighbor words as the context of each target word. Although it is an approximation, we can expect reliability improvement often seen by ignoring the lower part. In addition, this limitation of interactions highly reduces computational cost in particular when applying to larger-scale problems. To do this, the exhaustive sum $\sum_{i,\,i':\,w_i \neq w_{i'}}$ in Equation (7)-(9) is altered by the local sum $\sum_{i,\,i':\,(w_i, w_{i'}) \in P_{k\mathrm{NN}}}$, where $P_{k\mathrm{NN}}$ denotes the set of word pairs of which either is a k-nearest neighbors of the other. The normalizing factors 1, N, and $N - N_i$ in Equation (7), (8)-(9), and (11) are altered by the actual sum of responsibilities within those neighbors as $\sum_{i',\,j,\,j':\,(w_i, w_{i'}) \in P_{k\mathrm{NN}}} \mathcal{R}^{ij}_{i'j'}$,

$\sum_{i,\, i',\, j,\, j' :\, (w_i, w_{i'}) \in P_{k\mathrm{NN}}} \mathcal{R}^{ij}_{i'j'}$, and $\sum_{\iota,\, i',\, j,\, j' :\, (w_\iota, w_{i'}) \in P_{k\mathrm{NN}} \wedge \iota \neq i} \mathcal{R}^{\iota j}_{i'j'}$, respectively.

To treat the above similarity functions of context and of sense as distance functions, we use the conversion: $d(\cdot,\cdot) \equiv -\alpha \ln(f(\cdot,\cdot)/f_{\max})$, where d denotes the objective distance function, i.e., d_{x} for context and d_{s} for sense, while f and $f_{\max}$ denote the original similarity function and its maximum, respectively. α is a *standardization coefficient*, which is determined so that the mean squared distance be 1 in a dataset. According to this standardization, initial values of ${\sigma_{\mathrm{x}}}^2$, ${\sigma_{\mathrm{s}}}^2$ are always 1.

5 Evaluation

To confirm the effect of the proposed smoothing model and its combinatorial optimization scheme, we conducted WSD evaluations. The primary evaluations compare our method with conventional ones, in Section 5.2. Supplementary evaluations are described in the subsequent sections that include the comparison with SemEval-2 participating systems, and the analysis of model dynamics with the experimental data.

5.1 Evaluation Scheme

To make the evaluation comparable to state-of-the-art systems, we used the official dataset of the SemEval-2 English all-words WSD task (Agirre et al., 2010), which is currently the latest public dataset available with published results. The dataset consists of test data and background documents of the same *environment* domain. The test data consists of 1,398 target words (1,032 nouns and 366 verbs) in 5.3K running words. The background documents consists of 2.7M running words, which was used to compute distributional similarity.

Precisions and recalls were all computed using the official evaluation tool `scorer2` in fine-grained measure. The tool accepts answers either in probabilistic format (senses with probabilities for each target word) or in deterministic format (most likely senses, with no score information). As the proposed method is a probability model, we evaluated in the probabilistic way unless explicitly noted otherwise. For this reason, we evaluated all the sense probabilities as they were. Disambiguations were executed in separate runs for nouns and verbs, because no interaction takes place across POS in this metric implementation. The two runs' results were combined later to a single answer to be input to `scorer2`.

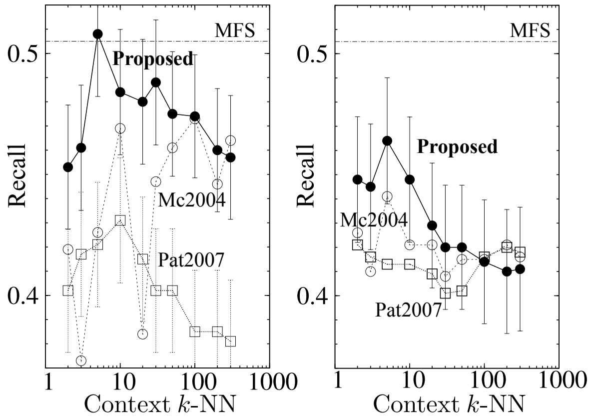
The context metric space was composed by k-nearest neighbor words of distributional similarity (Lin, 1998), as is described in Section 4. The value of k was evaluated for $\{2, 3, 5, 10, 20, 30, 50, 100, 200, 300\}$. As for sense metric space, we evaluated two measures i.e., (Jiang and Conrath, 1997) denoted as *JCN*, and (Lesk, 1986) denoted as *Lesk*. In every condition, stopping criterion of iteration is always the number of iteration (500 times), irrespective of the convergence in likelihood.

Primary evaluations compared our method with two conventional methods. Those methods differ to ours only in scoring schemes. The first one is the method by McCarthy et al. (2004), which determines the word sense based on sense similarity and distributional similarity to the k-nearest neighbor words of a target word by distributional similarity. Our major advantage is the combinatorial optimization framework, while the conventional one employs word-by-word scheme. The second one is based on the method by Patwardhan et al. (2007), which determines the word sense by maximizing the sum of sense similarity to the k immediate neighbor words of a target word. The k words were forced to be selected from other target words of the same POS to the word of interest, so as to make information resource equivalent to the other comparable two methods. It is also a word-by-word method. It exploits no distributional similarity. Our major advantages are the combinatorial optimization scheme and the smoothing model to integrate distributional similarity. In the following section, these comparative methods are referred to as *Mc2004* and *Pat2007*, respectively.

5.2 Comparison with Conventional Methods

Let us first confirm our advantages compared to the conventional methods of Mc2004 and Pat2007. The comparative results are shown in Figure 3 in recall measure. Precisions are simply omitted because the difference to the recalls are always the number of failures on referring to WordNet by mislabeling of lemmata or POSs, which is always the same for the three methods. Vertical range depicts 95% confidence intervals. The graphs also indicate the most-frequent-sense (MFS) baseline estimated from out-of-domain corpora, whose recall is 0.505 (Agirre et al., 2010).

As we can see in Figure 3(a) and 3(b), higher



(a) JCN (b) Lesk

Figure 3: Comparison to the conventional methods that differ to our method only in scoring schemes.

Table 1: Comparison with the top-5 knowledge-based systems in SemEval-2 (JCN/$k = 5$).

Rank	Participants	R	P	Rn	Rv
-	**Proposed (best)**	**50.8**	**51.0**	**52.5**	**46.2**
-	*MFS Baseline*	50.5	50.5	52.7	44.3
1	Kulkarni et al. (2010)	49.5	51.2	51.6	43.4
2	Tran et al. (2010)	49.3	50.6	51.6	42.6
3	Tran et al. (2010)	49.1	50.4	51.5	42.5
4	Soroa et al. (2010)	48.1	48.1	48.7	46.2
5	Tran et al. (2010)	47.9	49.2	49.4	43.4
...	...	...	...	...	...
-	*Random Baseline*	23.2	23.2	25.3	17.2

recalls are obtained in the order of the proposed method, Mc2004, and Pat2007 on the whole. Comparing JCN and Lesk, difference among the three is smaller in Lesk. It is possibly because Lesk is a score not normalized for different word pairs, which makes the effect of distributional similarity unsteady especially when combining many k-nearest words. Therefore the recalls are expected to improve if proper normalization is applied to the proposed method and Mc2004. In JCN, the recalls of the proposed method significantly improve compared to Pat2007. Our best recall is 0.508 with JCN and $k = 5$. Thus we can conclude that, though significance depends on metrics, our smoothing model and the optimization scheme are effective to improve accuracies.

5.3 Comparison with SemEval-2 Systems

We compared our best results with the participating systems of the task. Table 1 compares the details to the top-5 systems, which only includes unsupervised/knowledge-based ones and excludes supervised/weakly-supervised ones. Those values

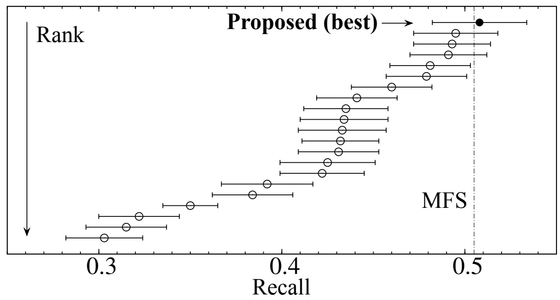


Figure 4: Comparison with the all 20 knowledge-based systems in SemEval-2 (JCN/$k = 5$).

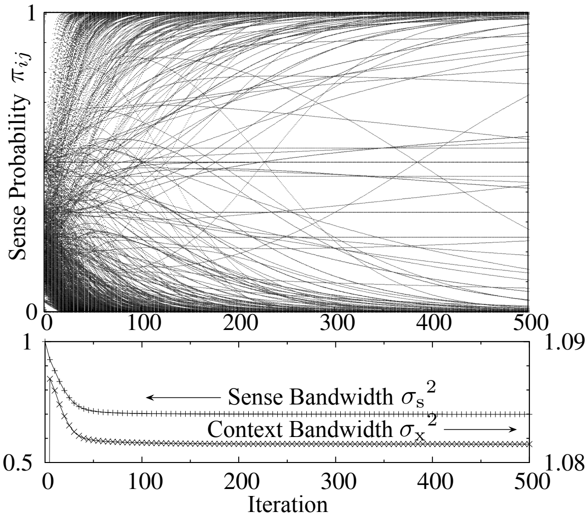


Figure 5: Model dynamics through iteration with SemEval-2 nouns (JCN/$k = 5$).

are transcribed from the official report (Agirre et al., 2010). "R" and "P" denote the recall and the precision for the whole dataset, while "Rn" and "Rv" denote the recall for nouns and verbs, respectively. The results are ranked by "R", in accordance with the original report. As shown in the table, our best results outperform all of the systems and the MFS baseline.

Overall rankings are depicted in Figure 4. It maps our best results in the distribution of all the 20 unsupervised/knowledge-based participating systems. The ranges spreading left and right are 95% confidence intervals. As is seen from the figure, our best results are located above the top group, which are outside the confidence intervals of the other participants ranked intermediate or lower.

5.4 Analysis on Model Dynamics

This section examines the model dynamics with the SemEval-2 data, which has been illustrated

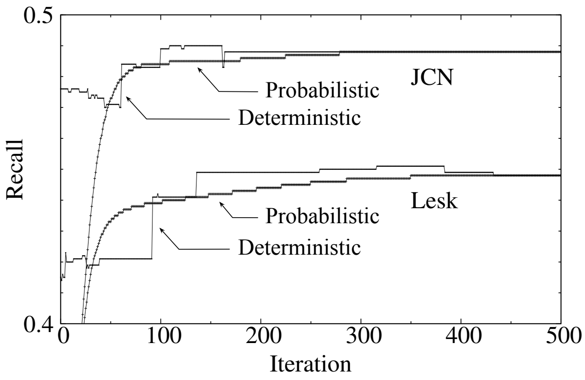


Figure 6: Recall improvement via iteration with SemEval-2 all POSs (JCN/k=30, Lesk/k=10).

with pseudo data in Section 3.2. Let us start by looking at the upper half of Figure 5, which shows the change of sense probabilities through iteration. At the initial status (iteration 0), the probabilities were all $1/2$ for bi-semous words, all $1/3$ for tri-semous words, and so forth. As iteration proceeded, the probabilities gradually spread out to either side of 1 or 0, and finally at iteration 500, we can observe that almost all the words were clearly disambiguated. The lower half of Figure 5 shows the dynamics in bandwidths. Vertical axis on the left is for the sense bandwidth, and on the right is for the context bandwidth. We can observe those bandwidths became narrower as iteration proceeded. Intensity of smoothing was dynamically adjusted by the whole disambiguation status. These behaviors confirm that even with an actual dataset, it works as is expected, just as illustrated in Figure 2.

6 Discussion

This section discusses the validity of the proposed method as to i) sense-interdependent disambiguation and ii) reliability of data smoothing. We here analyze the second peak conditions at $k = 30$ (JCN) and $k = 10$ (Lesk) instead of the first peak at $k = 5$, because we can observe tendency the better with the larger number of word interactions.

6.1 Effects of Sense-interdependent Disambiguation

Let us first examine the effect of our sense-interdependent disambiguation. We would like to confirm that how the progressive disambiguation is carried out. Figure 6 shows the change of recall through iteration for JCN ($k = 30$) and Lesk ($k = 10$). Those recalls were obtained by evaluating the status after each iteration. The recalls were here evaluated both in probabilistic format and in deterministic format. From the figure we can observe that the deterministic recalls also increased as well as the probabilistic recalls. This means that the ranks of sense candidates for each word were frequently altered through iteration, which further means that some new information not obtained earlier was delivered one after another to sense disambiguation for each word. From these results, we could confirm the expected sense-interdependency effect that a sense disambiguation of certain word affected to other words.

6.2 Reliability of Smoothing as Supervision

Let us now discuss the reliability of our smoothing model. In our method, sense disambiguation of a word is guided by its nearby words' extrapolation (smoothing). Sense accuracy fully depends on the reliability of the extrapolation. Generally speaking, statistical reliability increases as the number of random sampling increases. If we take sufficient number of *random words* as nearby words, the sense distribution comes close to the true distribution, and then we expect the statistically true sense distribution should find out the true sense of the target word, according to the *distributional hypotheses* (Harris, 1954). On the contrary, if we take nearby words that are biased to particular words, the sense distribution also becomes biased, and the extrapolation becomes less reliable.

We can compute the randomness of words that affect for sense disambiguation, by *word perplexity*. Let the word of interest be $w \in V$. The word perplexity is calculated as $2^{H|_w}$, where $H|_w$ denotes the entropy defined as $H|_w \equiv -\sum_{w' \in V \setminus \{w\}} p(w'|w) \log_2 p(w'|w)$. The conditional probability $p(w'|w)$ denotes the probability with which a certain word $w' \in V \setminus \{w\}$ determines the sense of w, which can be defined as the density ratio: $p(w'|w) \propto \sum_{i:\, w_i = w} \sum_{i':\, w_{i'} = w'} \sum_{j,j'} \mathcal{Q}_{i'j'}(h_{ij})$.

The relation between word perplexity and probability change for ground-truth senses of nouns (JCN/$k = 30$) is shown in Figure 7. The upper histogram shows the change in iteration 1-100, and the lower shows that of iteration 101-500. We divide the analysis at iteration 100, because roughly until the 100th iteration, the change in bandwidths converged, and the number of words to interact settled, as can be seen in Figure 5. The bars that

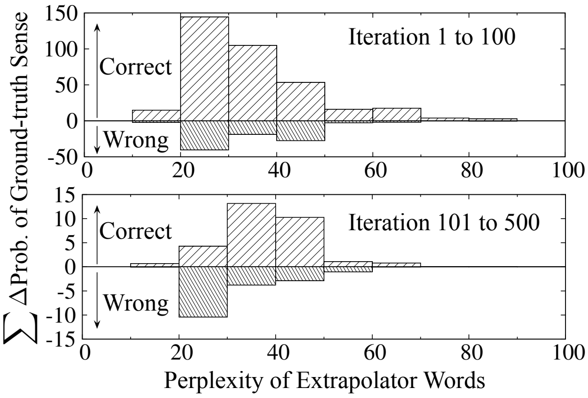


Figure 7: Correlation between reliability and perplexity with SemEval-2 nouns (JCN/$k = 30$).

extend upward represent the sum of the amount raised (correct change), and the bars that extend downward represent the sum of the amount reduced (wrong change). From these figures, we observe that when perplexity is sufficiently large (≥ 30), change occurred largely (79%) to the correct direction. In contrast, at the lower left of the figure, where perplexity is small (< 30) and bandwidths has been narrowed at iteration 101-500, correct change occupied only 32% of the whole. Therefore, we can conclude that if sufficiently random samples of nearby words are provided, our smoothing model is reliable, though it is trained in an unsupervised fashion.

7 Related Work

As described in Section 1, graph-based WSD has been extensively studied, since graphs are favorable structure to deal with interactions of data on vertices. Conventional studies typically consider as vertices the instances of input or target class, e.g. knowledge-based approaches typically regard *senses as vertices* (see Section 1), and corpus-based approaches such as (Véronis, 2004) regard *words as vertices* or (Niu et al., 2005) regards *context as vertices*. Our method can be viewed as one of graph-based methods, but it regards *input-to-class mapping as vertices*, and the edges represent the relations both together in context and in sense. Mihalcea (2005) proposed graph-based methods, whose vertices are sense label hypotheses on word sequence. Our method generalize context representation.

In the evaluation, our method was compared to SemEval-2 systems. The main subject of the SemEval-2 task was domain adaptation, therefore those systems each exploited their own adaptation techniques. Kulkarni et al. (2010) used a WordNet pre-pruning. Disambiguation is performed by considering only those candidate synsets that belong to the top-k largest connected components of the WordNet on domain corpus. Tran et al. (2010) used over 3TB domain documents acquired by Web search. They parsed those documents and extracted the statistics on dependency relation for disambiguation. Soroa et al. (2010) used the method by Agirre et al. (2009) described in Section 1. They disambiguated each target word using its distributionally similar words instead of its immediate context words.

The proposed method is an extension of density estimation (Parzen, 1962), which is a construction of an estimate based on *observed data*. Our method naturally extends the density estimation in two points, which make it applicable to unsupervised knowledge-based WSD. First, we introduce stochastic treatment of data, which is no longer observations but hypotheses having ambiguity. This extension makes the hypotheses possible to cross-validate the plausibility each other. Second, we extend the definition of density from Euclidean distance to general metric, which makes the proposed method applicable to a wide variety of corpus-based context similarities and dictionary-based sense similarities.

8 Conclusions

We proposed a novel *smoothing model* with a *combinatorial optimization scheme* for all-words WSD from untagged corpora. Experimental results showed that our method significantly improves the accuracy of conventional methods by exceeding most-frequent-sense baseline performance where none of SemEval-2 unsupervised systems reached. Detailed inspection of dynamics clearly show that the proposed optimization method effectively exploit the sense-dependency of all-words. Moreover, our smoothing model, though unsupervised, provides reliable supervision when sufficiently random samples of words are available as nearby words. Thus it was confirmed that this method is valid for finding the optimal combination of word senses with large untagged corpora. We hope this study would elicit further investigation in this important area.

References

Eneko Agirre and Philip Edmonds. 2006. *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science+ Business Media.

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41.

Eneko Agirre, Oier Lopez De Lacalle, Aitor Soroa, and Informatika Fakultatea. 2009. Knowledge-based wsd on specific domains: performing better than generic supervised wsd. In *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 1501–1506.

Eneko Agirre, Oier Lopez de Lacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. 2010. Semeval-2010 task 17: All-words word sense disambiguation on a specific domain. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 75–80.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the rasp system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 77–80.

Arthur Pentland Dempster, Nan McKenzie Laird, and Donald Bruce Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

Zellig Sabbetai Harris. 1954. Distributional structure. *Word*.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Anup Kulkarni, Mitesh M. Khapra, Saurabh Sohoney, and Pushpak Bhattacharyya. 2010. CFILT: Resource conscious approaches for all-words domain specific. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 421–426.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 279–286.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590.

Rada Mihalcea. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 411–418.

Roberto Navigli and Mirella Lapata. 2007. Graph connectivity measures for unsupervised word sense disambiguation. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1683–1688.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Zheng-Yu Niu, Dong-Hong Ji, and Chew Lim Tan. 2005. Word sense disambiguation using label propagation based semi-supervised learning. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 395–402.

Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2007. UMND1: Unsupervised word sense disambiguation using contextual semantic relatedness. In *proceedings of the 4th International Workshop on Semantic Evaluations*, pages 390–393.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, pages 38–41.

Aitor Soroa, Eneko Agirre, Oier Lopez de Lacalle, Monica Monachini, Jessie Lo, Shu-Kai Hsieh, Wauter Bosma, and Piek Vossen. 2010. Kyoto: An integrated system for specific domain WSD. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 417–420.

Andrew Tran, Chris Bowes, David Brown, Ping Chen, Max Choly, and Wei Ding. 2010. TreeMatch: A fully unsupervised WSD system using dependency knowledge on a specific domain. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 396–401.

Jean Véronis. 2004. HyperLex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252.