An Evidential Reasoning Approach to Weighted Combination of Classifiers for Word Sense Disambiguation

Cuong Anh Le[1], Van-Nam Huynh[2], and Akira Shimazu[1]

[1] School of Information Science
[2] School of Knowledge Science
Japan Advanced Institute of Science and Technology
Tatsunokuchi, Ishikawa, 923-1292, JAPAN
Email: {cuonganh,huynh,shimazu}@jaist.ac.jp

Abstract. Arguing that various ways of using context in word sense disambiguation (WSD) can be considered as distinct representations of a polysemous word, a theoretical framework for the weighted combination of soft decisions generated by experts employing these distinct representations is proposed in this paper. Essentially, this approach is based on the Dempster-Shafer theory of evidence. By taking the confidence of individual classifiers into account, a general rule of weighted combination for classifiers is formulated, and then two particular combination schemes are derived. These proposed strategies are experimentally tested on the datasets for four polysemous words, namely *interest*, *line*, *serve*, and *hard*, and obtained the better result in comparison with previous studies for all cases, with an exception of the word *line*.

Keywords: Computational linguistics, Weighted combination of classifiers, Word sense disambiguation, Dempster-Shafer theory of evidence.

1 Introduction

Word sense disambiguation is a computational linguistics task recognized since the 1950s. Roughly speaking, word sense disambiguation involves the association of a given word in a text or discourse with a particular sense among numerous potential senses of that word. As mentioned in [5], this is an "intermediate task" necessarily to accomplish most natural language processing tasks. It is obviously essential for language understanding applications, while also at least helpful for other applications whose aim is not language understanding such as machine translation, information retrieval, among others. Since its inception, many methods involving WSD have been developed in the literature (see, e.g., [5] for a survey). During the last decade, many supervised machine learning algorithms have been used for this task, including Naïve Bayesian (NB) model, decision trees, exemplar-based model, support vector machine, maximum entropy, etc. As observed in studies of machine learning systems, although one could choose

one of learning systems available to achieve the best performance for a given pattern recognition problem, the set of patterns misclassified by the different classification systems would not necessarily overlap. This means that different classifiers may potentially offer complementary information about the patterns to be classified. This observation highly motivated the interest in combining classifiers during the recent years. Especially, classifier combination for WSD has been unsurprisingly received much attention recently from the community as well, e.g., [6, 4, 13, 8, 3, 16].

As is well-known, there are basically two classifier combination scenarios. In the first scenario, all classifiers use the same representation of the input pattern. In the context of WSD, the work by Kilgarriff and Rosenxweig [6], Klein et al. [8], and Florian and Yarowsky [3] could be grouped into this first scenario. In the second scenario, each classifier uses its own representation of the input pattern. An important application of combining classifiers in this scenario is the possibility to integrate physically different types of features. In this sense, the work by Pedersen [13] can be considered as belonging to this scenario, although the difference of representations here is only in terms of size of context windows. Also, Wang and Matsumoto [16] used similar sets of features as in Pedersen [13], but proposed a new strategy of voting based on kNN method.

An important issue in combining classifiers is the combination strategy used to derive a consensus decision. In this paper, we focus on the weighted combination of classifiers for WSD in the second scenario mentioned above. Particularly, we first consider various ways of using context in WSD as distinct representations of a polysemous word under consideration, then all these representations are used as providing individual information sources to identify the meaning of the target word. We then develop a general framework for the weighted combination of individual classifiers corresponding to distinct representations. Essentially, this approach is based on Dempster-Shafer (DS) theory of evidence [14], which has been recently increasingly applied to classification problems, e.g. [2, 17]. Moreover, two combination strategies are developed and experimentally tested on the datasets for four polysemous words, namely *interest*, *line*, *serve*, and *hard*, and compared with previous studies.

The paper is organized as follows. In the next section, basic notions of DS theory will be briefly recalled. Section 3 necessarily reformulate the WSD problem so that the general framework for the weighted combination of classifiers can be formulated, and the two combination strategies can be developed. In Section 3, we describe a multi-representation scheme for context in WSD problem. Section 4 first proposes an effective computation of necessary probabilities, and then presents experimented results and some comparison with previous known results on the same test datasets. Finally, some conclusions are presented in Section 5.

2 Dempster-Shafer Theory of Evidence

In DS theory, a problem domain is represented by a finite set Θ of mutually exclusive and exhaustive hypotheses, called *frame of discernment* [14]. In the standard probability framework, all elements in Θ are assigned a probability. And when the degree of support for an event is known, the remainder of the support is automatically assigned to the negation of the event. On the other hand, in DS theory mass assignments are carried out for events as they know, and committing support for an event does not necessarily imply that the remaining support is committed to its negation. Formally, a basic probability assignment (BPA, for short) is a function $m : 2^{\Theta} \rightarrow [0, 1]$ verifying

$$m(\emptyset) = 0, \text{ and } \sum_{A \in 2^{\Theta}} m(A) = 1$$

The quantity $m(A)$ can be interpreted as a measure of the belief that is committed exactly to A, given the available evidence. A subset $A \in 2^{\Theta}$ with $m(A) > 0$ is called a *focal element* of m. A BPA m is called to be *vacuous* if $m(\Theta) = 1$ and $m(A) = 0$ for all $A \neq \Theta$.

Two evidential functions derived from the basic probability assignment m are the belief function Bel_m and the plausibility function Pl_m, defined as

$$Bel_m(A) = \sum_{\emptyset \neq B \subseteq A} m(B), \text{ and } Pl_m(A) = \sum_{B \cap A \neq \emptyset} m(B)$$

The difference between $m(A)$ and $Bel_m(A)$ is that while $m(A)$ is our belief committed to the subset A excluding any of its proper subsets, $Bel_m(A)$ is our degree of belief in A as well as all of its subsets. Consequently, $Pl_m(A)$ represents the degree to which the evidence fails to refute A. Note that all the three functions are in an one-to-one correspondence with each other.

Two useful operations that play a central role in the manipulation of belief functions are *discounting* and *Dempster's rule of combination* [14]. The discounting operation is used when a source of information provides a BPA m, but one knows that this source has probability α of reliable. Then one may adopt $(1-\alpha)$ as one's *discount rate*, which results in a new BPA m^{α} defined by

$$m^{\alpha}(A) = \alpha m(A), \quad \text{for any } A \subset \Theta \tag{1}$$

$$m^{\alpha}(\Theta) = (1 - \alpha) + \alpha m(\Theta) \tag{2}$$

Consider now two pieces of evidence on the same frame Θ represented by two BPAs m_1 and m_2. Dempster's rule of combination is then used to generate a new BPA, denoted by $(m_1 \oplus m_2)$ (also called the orthogonal sum of m_1 and m_2), defined as follows

$$\begin{aligned} (m_1 \oplus m_2)(\emptyset) &= 0, \\ (m_1 \oplus m_2)(A) &= \tfrac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C) \end{aligned} \tag{3}$$

where

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{4}$$

Note that the orthogonal sum combination is only applicable to such two BPAs that verify the condition $\kappa < 1$.

It is worth noting that Dempster rule of combination has some attractive features such as: it is commutative and associative; given two BPAs m_1 and m_2, if m_1 is vacuous then $m_1 \oplus m_2 = m_2$.

3 Weighted Combination of Classifiers for WSD

In this section, after reformulating the WSD problem in terms of a pattern recognition problem with multi-representation of patterns. The general framework for weighted combination of classifiers is developed for WSD problem and then, two particular combination schemes are explored.

3.1 WSD with Multi-Representation of Context

As is well-known, in WSD problem, context plays an essentially important role to identify the meaning of a polysemous work. Given an polysemous word w, which may have M possible senses (classes): c_1, c_2,..., c_M, in a context C, the task is to determine the most appropriate sense of w.

Generally, context C can be used in two ways [5]: in the *bag-of-words approach*, the context is considered as words in some window surrounding the target word w; in the *relational information based approach*, the context is considered in terms of some relation to the target such as distance from the target, syntactic relations, selectional preferences, phrasal collocation, semantic categories, etc. As such, for a target word w, we may have different representations of context C corresponding to different views of context. Assume we have such R representations of C, say $\mathbf{f}_1, \ldots, \mathbf{f}_R$, serving for the aim of identifying the right sense of the target w. Clearly, each $\mathbf{f}_i$ can be also considered as a semantical representation of w. Each representation $\mathbf{f}_i$ of context has its own type depending on which way context is used (for the detail, see Section 4). In the sequent, we can use a set of features and a representation interchangeably without danger of confusion.

Now let us assume that we have R classifiers, each representing the context by a distinct set of features. The set of features $\mathbf{f}_i$, which is considered as a representation of context C of the target w, is used by the i-th classifier. Due to the interpretation of $\mathbf{f}_i$'s and the role of context in WSD, quite naturally, we shall assume that the individual models corresponding to different representations of context are independent. Furthermore, assume that each i-th classifier (expert) is associated with a weight α_i, $0 \leq \alpha_i \leq 1$, reflecting the relative confidence in it, which may be interpreted as reliable probability of the i-th classifier in its prediction.

As such representations $\mathbf{f}_i$'s ($i = 1, \ldots, R$) are considered as distinct information sources associated with corresponding weights serving for identifying the sense of the target w. The problem now is how to combine these information sources to reach a consensus decision for identifying the sense of w.

3.2 A General Framework

Given a target word w in a context C and $\mathcal{S} = \{c_1, c_2, \ldots, c_M\}$ is the set of its possible senses. Using the vocabulary of DS theory, $\mathcal{S}$ can be called the *frame of discernment* of the problem. As mentioned above, various ways of using the context could be considered as providing different information sources to identify the meaning of the target word. Each of these information sources does not by itself provide 100% certainty as a whole piece of evidence for identifying the sense of the target. Formally, we have the available information for making the final decision on the sense of w given as follows

- R probability distributions $P(\cdot|\mathbf{f}_i)$ ($i = 1, \ldots, R$) on $\mathcal{S}$,
- the weights α_i of the individual information sources ($i = 1, \ldots, R$)[3].

From the probabilistic point of view, we may straightforwardly think of the combiner as a weighted mixture of individual classifiers defined as

$$P(c_k) = \frac{1}{\sum_i \alpha_i} \sum_{i=1}^{R} \alpha_i P(c_k|\mathbf{f}_i), \text{ for } k = 1, \ldots, M \tag{5}$$

Then the target word w should be naturally assigned to the sense c_j according to the following decision rule

$$j = \arg\max_k P(c_k) \tag{6}$$

However, by considering the problem as that of weighted combination of evidence for decision making, in the following we will formulate a general rule of combination based on DS theory. To this end, we first adopt a probabilistic interpretation of weights. That is, the weight α_i ($i = 1, \ldots, R$) is interpreted as reliable probability of the i-th classifier. This interpretation of weights seems to be especially appropriate when defining weights in terms of the accuracy of individual classifiers.

Under such an interpretation of weights, the piece of evidence represented by $P(\cdot|\mathbf{f}_i)$ should be discounted at a discount rate of $(1 - \alpha_i)$. This results in a BPA m_i verifying

$$m_i(\{c_k\}) = \alpha_i P(c_k|\mathbf{f}_i) \triangleq p_{i,k}, \text{ for } k = 1, \ldots, M \tag{7}$$

$$m_i(\mathcal{S}) = 1 - \alpha_i \triangleq p_{i,\mathcal{S}} \tag{8}$$

$$m_i(A) = 0, \forall A \in 2^{\mathcal{S}} \setminus \{\mathcal{S}, \{c_1\}, \ldots, \{c_M\}\} \tag{9}$$

[3] Note that the constraint $\sum_i \alpha_i = 1$ does not need to be imposed.

That is, the discount rate of $(1-\alpha_i)$ can not be distributed to anything else than $\mathcal{S}$, the whole frame of discernment.

We are now ready to formulate our belief on the decision problem by aggregating all pieces of evidence represented by m_i's in the general form of the following

$$m = \bigoplus_{i=1}^{R} m_i \tag{10}$$

where m is a BPA and $\oplus$ is a combination operator in general.

As such, by applying different combination operations for $\oplus$, we may have different aggregation schemes for obtaining the BPA m which modelled our belief for making the decision on the sense of w. Therefore, we must also deal with the problem of how to make a decision based on m. As m does not in general provide a unique probability distribution on $\mathcal{S}$, but only a set of *compatible probabilities* bounded by the belief function Bel_m and the plausibility function Pl_m. Consequently, individual classes in $\mathcal{S}$ can no longer be ranked according to their probability. Fortunately, based on the *Generalized Insufficient Reason Principle*, we may define a probability function P_m on $\mathcal{S}$ derived from m for the purpose of decision making via the *pignistic transformation* [15]. That is, as in the two-level language of the so-called *transferable belief model* [15], the aggregated BPA m itself represented the belief is entertained based on the available evidence at the *credal level*, and when a decision must be made, the belief at the credal level induces the probability function P_m for decision making.

3.3 The Discounting-and-Orthogonal Sum Combination Strategy

As discussed above, we consider each $P(\cdot|\mathbf{f}_i)$ as the belief quantified from the information source $\mathbf{f}_i$ and the weight α_i as a "degree of trust" of $\mathbf{f}_i$ supporting the identification for the sense of w as a whole. As mentioned in [14], an obvious way to use discounting with Dempster's rule of combination is to discount all BPAs $P(\cdot|\mathbf{f}_i)$ $(i=1,\ldots,R)$ at corresponding rates $(1-\alpha_i)$ $(i=1,\ldots,R)$ before combining them.

Thus, Dempster's rule of combination now allows us to combine BPAs m_i $(i=1,\ldots,R)$ under the independent assumption of information sources for generating the BPA m, i.e. $\oplus$ in (10) is the orthogonal sum operation.

Note that, by definition, focal elements of each m_i are either singleton sets or the whole set $\mathcal{S}$. It is easy to see that m also verifies this property if applicable. Interestingly, the commutative and associative properties of the orthogonal sum operation with respect to a combinable collection of BPAs m_i $(i=1,\ldots,M)$ and the mentioned property essentially form the basis for developing a recursive algorithm for calculation of the BPA m. This can be done as follows.

Let $I(i)=\{1,\ldots,i\}$ be the subset consisting of first i indexes of the set $\{1,\ldots,R\}$. Assume that $m_{I(i)}$ is the result of combining the first i BPAs m_j, for

$j = 1, \ldots, i$. Let us denote

$$p_{I(i),k} \triangleq m_{I(i)}(\{c_k\}), \text{ for } k = 1, \ldots, M \tag{11}$$

$$p_{I(i),\mathcal{S}} \triangleq m_{I(i)}(\mathcal{S}) \tag{12}$$

With these notations and (7)–(8), the key step in the combination algorithm is to inductively calculate $p_{I(i+1),k}$ $(k = 1, \ldots, M)$ and $p_{I(i+1),\mathcal{S}}$ as follows

$$p_{I(i+1),k} = \frac{1}{\kappa_{I(i+1)}}[p_{I(i),k}p_{i+1,k} + p_{I(i),k}p_{i+1,\mathcal{S}} + p_{I(i),\mathcal{S}}p_{i+1,k}] \tag{13}$$

$$p_{I(i+1),\mathcal{S}} = \frac{1}{\kappa_{I(i+1)}}(p_{I(i),\mathcal{S}}p_{i+1,\mathcal{S}}) \tag{14}$$

for $k = 1, \ldots, M$, $i = 1, \ldots, R-1$, and $\kappa_{I(i+1)}$ is a normalizing factor defined by

$$\kappa_{I(i+1)} = \left[1 - \sum_{j=1}^{M} \sum_{\substack{k=1 \\ k \neq j}}^{M} p_{I(i),j}p_{i+1,k}\right] \tag{15}$$

Finally, we obtain m as $m_{I(R)}$. For the purpose of decision making, we now define a probability function P_m on $\mathcal{S}$ derived from m via the *pignistic transformation* as follows

$$P_m(c_k) = m(\{c_k\}) + \frac{1}{M}m(\mathcal{S}) \text{ for } k = 1, \ldots, M \tag{16}$$

and we have the following decision rule:

$$j = \arg\max_k P_m(c_k) \tag{17}$$

It would be interesting to note that an issue may arise with the orthogonal sum operation, that is the use of the total probability mass κ associated with conflict as defined in the normalization factor. Consequently, applying it in an aggregation process may yield counterintuitive results in the face of significant conflict in certain situations as pointed out in [18]. Fortunately, in the context of the weighted combination of classifiers, by discounting all $P(\cdot|\mathbf{f}_i$ $(i = 1, \ldots, R)$ at corresponding rates $(1 - \alpha_i)$ $(i = 1, \ldots, R)$, we actually reduce conflict between the individual classifiers before combining them.

3.4 The Discounting-and-Averaging Combination Strategy

In this strategy, instead of using Dempster's rule of combination after discounting $P(\cdot|\mathbf{f}_i)$ at the discount rate of $(1 - \alpha_i)$, we apply the averaging operation over BPAs m_i $(i = 1, \ldots, R)$ to obtain the BPA m defined by

$$m(A) = \frac{1}{R}\sum_{i=1}^{R} m_i(A) \tag{18}$$

for any $A \in 2^{\mathcal{S}}$. By definition, we get

$$m(\{c_k\}) = \frac{1}{R}\sum_{i=1}^{R} \alpha_i P(c_k|\mathbf{f}_i), \text{ for } k = 1, \ldots, M \tag{19}$$

$$m(\mathcal{S}) = 1 - \frac{\sum_{i=1}^{R} \alpha_i}{R} \triangleq 1 - \overline{\alpha} \tag{20}$$

$$m(A) = 0, \forall A \in 2^{\mathcal{S}} \setminus \{\mathcal{S}, \{c_1\}, \ldots, \{c_M\}\} \tag{21}$$

Note that the probability mass unassigned to individual classes but the whole frame of discernment $\mathcal{S}$, $m(\mathcal{S})$, is the average of discount rates. Therefore, if instead of allocating the average discount rate $(1 - \overline{\alpha})$ to $m(\mathcal{S})$ as above, we use it as a normalization factor and easily obtain

$$m(\{c_k\}) = \frac{1}{\sum_i \alpha_i}\sum_{i=1}^{R} \alpha_i P(c_k|\mathbf{f}_i), \text{ for } k = 1, \ldots, M \tag{22}$$

$$m(A) = 0, \forall A \in 2^{\mathcal{S}} \setminus \{\{c_1\}, \ldots, \{c_M\}\} \tag{23}$$

which interestingly turns out to be the weighted mixture of individual classifiers as defined in (5). Then we have the decision rule (6).

It should be worth noting that since the average discount rate $(1 - \overline{\alpha})$ is a constant, the decision rule based on the weighted mixture of individual classifiers is the same as that based on the probability function P_m with m is defined by (19)–(21) via the pignistic transformation.

In the following we will experimentally test the proposed combination strategies on the datasets for four polysemous words, namely *interest*, *line*, *serve*, and *hard*, and also compare the experimental results with previous studies.

4 Representations of Context for WSD

As mentioned previously, context plays an essentially important role in WSD, therefore, the representation choice of context is a factor which may be more important than the algorithm used for the task itself on the aspect of affecting the obtained result. For predicting senses of a word, information usually used in all studies is the topic context which is represented by bag of words. In [12], Ng and Lee proposed a use of more linguistic knowledge resources, which then became popular resources for determining word sense in many papers. The knowledge resources used in their paper included topic context, collocation of words, and a syntactic relationship verb-object. In [10], the authors use another information type, which is words or part-of-speech and each is assigned with its position in relation with the target word. However, in the second scenario of classifier combination, according to our knowledge, only topic context with different sizes of context windows is used for creating different representations of a polysemous word, such as in Pedersen [13] and Wang and Matsumoto [16].

It is worth to emphasize that, two of the most important kinds of information for determining the sense of a polysemous word are the topic of the context and relational information representing the structural relations between the target word and the surrounding words in a local context. A bag of unordered words in the context can determine the topic of the context and collocation can determine grammatical information. Ordered words in a local context are also an important resource for relational information. We did not use syntactical relations such as verb-object, which are used by Ng and Lee in [12], because this information can be found in collocation features and a syntactic parser does not always output a correct result. In this work we use five kinds representations corresponding to five classifiers each represents a set of features as proposed in Le and Shimazu [9], as following:

- $\mathbf{f}_1$ is a set of unordered words in the large context, namely

$$\mathbf{f}_1 = \{w_{-n_1}, \ldots, w_{-2}, w_{-1}, w_1, w_2, \ldots, w_{n_1}\}$$

- $\mathbf{f}_2$ is a set of words assigned with their positions in the local context, namely

$$\mathbf{f}_2 = \{(w_{-n_2}, -n_2), \ldots, (w_{-2}, -2), (w_{-1}, -1), (w_1, 1), (w_2, 2), \ldots, (w_{n_2}, n_2)\}$$

- $\mathbf{f}_3$ is a set of part-of-speech tags assigned with their positions in the local context, namely

$$\mathbf{f}_3 = \{(p_{-n_3}, -n_3), \ldots, (p_{-2}, -2), (p_{-1}, -1), (p_1, 1), (p_2, 2), \ldots, (p_{n_3}, n_3)\}$$

- $\mathbf{f}_4$ is a set of collocations of words, namely

$$\mathbf{f}_4 = \{w_{-l} \cdots w_{-1} w w_1 \cdots w_r | l + r \leq n_4\}$$

- $\mathbf{f}_5$ is a set of collocations of part-of-speech tags, namely

$$\mathbf{f}_5 = \{p_{-l} \cdots p_{-1} w p_1 \cdots p_r |\ l + r \leq n_5\}$$

Where w_i is the word at position i in the context of the ambiguous word w and p_i be the part-of-speech tag of w_i, with the convention that the target word w appears precisely at position 0 and i will be negative (positive) if w_i appears on the left (right) of w.

In the experiment, we design the window size of topic context (for both left and right windows) as 50 for the representation $\mathbf{f}_1$, i.e. $n_1 = 50$, while the window size of local context as 3 for remaining representations, i.e. $n_i = 3$, for $i = 2, 3, 4, 5$. Our representations for the individual classifiers are richer than the representation that just used the words in context because we also use the feature containing richer information about structural relations. Even that the unordered words in a local context may also contain structure information, but collocations and words and part-of-speech tags assigned with their positions of course will bring richer information.

5 Experiments

5.1 Data

We tested on the datasets for four words, namely *interest*, *line*, *serve*, and *hard*, which are used in numerous comparative studies of word sense disambiguation methodologies such as Pedersen [13], Ng and Lee [12], Bruce & Wiebe [1], and Leacock and Chodorow [10]. We have obtained those data from Pedersen's homepage [4]. There are 2369 instances of *interest* with 6 senses, 4143 instances of *line* with 6 senses, 4378 instances of *serve* with 4 senses, and 4342 instances of *hard* with 3 senses.

5.2 Computing the probabilities and determining weights

As obviously seen above, in the weighted combination of classifiers we need to compute the a posteriori probabilities $P(c_k|\mathbf{f}_i)$. For the context C, suppose that the representation $\mathbf{f}_i$ of C is represented by a set of features $\mathbf{f}_i = (f_{i,1}, f_{i,2}, \ldots, f_{i,n_i})$, and that the features $f_{i,j}$ are conditionally independent, we have:

$$P(c_k|\mathbf{f}_i) = \frac{P(\mathbf{f}_i|c_k)P(c_k)}{P(\mathbf{f}_i)} = \frac{P(c_k)\prod_{j=1}^{n_i} P(f_{i,j}|c_k)}{P(\mathbf{f}_i)} \tag{24}$$

However, due to the nature of multiplication with small numbers, if we compute $P(c_k|\mathbf{f}_i)$ directly, it may cause undesirable errors which strongly affects the final results, for example, the obtained value may be much far from the true value because the product of $P(f_{i,j}|c_k)$, $j = 1, \ldots, n_i$, may be too small. To avoid such an effect during the computation, in the following we provide an indirectly alternative way to compute $P(c_k|\mathbf{f}_i)$ more exactly. This can be done as follows.

For simplicity, assume that we are working on the representation $\mathbf{f}_i$, we then have

$$\sum_{k=1}^{m} P(c_k|\mathbf{f}_i) = 1$$

Let us denote

$$r_k = \frac{P(c_k|\mathbf{f}_i)}{P(c_1|\mathbf{f}_i)}, \text{ for } k = 1, \ldots, M$$

With this notation, we immediately obtain

$$P(c_1|\mathbf{f}_i) = \frac{1}{\sum_{k=1}^{m} r_k} \tag{25}$$

Clearly, $r_1 = 1$. We will then compute r_k $(k = 2, \ldots, M)$ based on the following formulation. From (24), we have

$$r_k = \frac{P(c_k|\mathbf{f}_i)}{P(c_1|\mathbf{f}_i)} = \frac{P(c_k)\prod_{j=1}^{n_i} P(f_{i,j}|c_k)}{P(c_1)\prod_{j=1}^{n_i} P(f_{i,j}|c_1)}$$

[4] http://www.d.umn.edu/∼tpederse/data.html

Logarithmizing the last expression we obtain

$$\log(r_k) = \sum_{j=1}^{n_i} \log(P(f_{i,j}|c_k)) + \log(P(c_k)) - \sum_{j=1}^{n_i} \log(P(f_{i,j}|c_1)) - \log(P(c_1)) \quad (26)$$

which is easy to compute more exactly. Once all r_k are computed via (26), it is easily to derive probabilities $P(c_k|\mathbf{f}_i)$, for $k = 1, \ldots, M$, from (25).

The probability of sense c_k, $P(c_k)$, and the conditional probability of a feature $f_{i,j}$ with observation of sense c_k, $P(f_{i,j}|c_k)$, are computed via maximum-likelihood estimation as:

$$P(c_k) = \frac{\mathsf{count}(c_k)}{N}$$

and

$$P(f_{i,j}|w = c_k) = \frac{\mathsf{count}(f_{i,j}, c_k)}{\mathsf{count}(c_k)}$$

where $\mathsf{count}(f_{i,j}, c_k)$ is the number of occurrences of $f_{i,j}$ in a context of sense c_k in the training corpus, $\mathsf{count}(c_k)$ is the number of occurrences of c_k in the training corpus, and N is the total number of occurrences of the polysemous word w or the size of the training dataset. To avoid the effects of zero counts when estimating the conditional probabilities of the model, when meeting a new feature $f_{i,j}$ in a context of the test dataset, for each sense c_k we set $P(f_{i,j}|w = c_k)$ equal to $\frac{1}{N}$.

In the experiment, we used 10-fold cross validation on the training data and then the obtained accuracies of the individual classifiers are used for weights α_i. Although we determine the weights based on the accuracies of individual classifiers, other methods of identifying the weights α_i such as using linear regression and least-squares-fit could be used. However, this is left for the long version of this paper.

5.3 Result and Comparison

In the experiment, we obtained the results using 10-fold cross validation. Data included four datasets corresponding to four polysemous words *interest*, *line*, *hard*, and *serve*. Table 1 shows the results obtained by using two strategies of weighted combination of classifiers and the best results obtained by individual classifiers respectively. It is shown that both combination strategies give better results than the best individual classifier in all cases. Interestingly also, the results showed that in all cases the orthogonal sum based combination strategy is better than that based on weighted sum. This can be experimentally interpreted as follows. In our multi-representation of context, each individual classifier corresponds to a type of features so that the conditional independence assumption seems to be realistic and, consequently, the orthogonal sum based combination strategy is a suitable choice for this scheme of multi-representation of context.

In Table 2, we show the obtained results in comparison with those taken from previous studies, which were only tested on several of these four words. It

Table 1. Results using the proposed method

	Best individual classifier (%)	Orthogonal sum combiner (%)	Weighted sum combiner (%)
interest	86.8	**90.9**	90.7
line	82.8	**87.2**	85.6
hard	90.2	**91.5**	91
serve	84.4	**89.7**	89

is shown that both combination strategies also give better results than previous methods in all cases, with the exception of *line* which corresponds to Pedersen's method as the best.

Table 2. The comparison with previous studies

(%)	BW[5]	M	NL	LC	P	The proposed method	
						based on weighted sum	based on orthogonal sum
interest	78	–	87	–	89	90.7	**90.9**
line	–	72	–	84	**88**	85.6	87.2
hard	–	–	–	83	–	91	**91.5**
serve	–	–	–	83	–	89	**89.7**

6 Conclusion

In this paper we first argued that various ways of using context in WSD can be considered as distinct representations of a polysemous word under consideration, then these representations are used jointly with taking weights into account to identify the meaning of the target word. Based on DS theory of evidence, we developed a general framework for the weighted combination of individual classifiers corresponding to distinct representations. Moreover, two combination strategies have been developed and experimentally tested on the datasets for four polysemous words, namely *interest*, *line*, *serve*, and *hard*, and compared with previous studies. It has been shown that considering multi-representation of context significantly improves the accuracy of WSD by combining classifiers, as individual classifiers corresponding to different types of representation suitably

offer complementary information about the target to be assigned a sense, this consequently helps to make more correct decisions.

Acknowledgement

This research is partly conducted as a program for the "Fostering Talent in Emergent Research Fields" in Special Coordination Funds for Promoting Science and Technology by the Japanese Ministry of Education, Culture, Sports, Science and Technology.

References

1. Bruce, R. and Wiebe, J. 1994. Word-Sense Disambiguation using Decomposable Models. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 139–145.
2. Denoeux, T., A k-nearest neighbor classification rule based on Dempster-Shafer theory, *IEEE Transactions on Systems, Man and Cybernetics* **25** (05) (1995) 804–813.
3. Florian, R., and D. Yarowsky, Modeling consensus: Classifier combination for Word Sense Disambiguation, *Proceedings of EMNLP 2002*, pp. 25–32.
4. Hoste, V., I. Hendrickx, W. Daelemans, and A. van den Bosch, Parameter optimization for machine-learning of word sense disambiguation, *Natural Language Engineering* **8** (3) (2002) 311–325.
5. Ide, N., J. Véronis, Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art, *Computational Linguistics* **24** (1998) 1–40.
6. Kilgarriff, A., and J. Rosenzweig, Framework and results for English SENSEVAL, *Computers and the Humanities* **36** (2000) 15–48.
7. Kittler, J., M. Hatef, R. P. W. Duin, and J. Matas, On combining classifiers, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** (3) (1998) 226–239.
8. Klein, D., K. Toutanova, H. Tolga Ilhan, S. D. Kamvar, and C. D. Manning, Combining heterogeneous classifiers for Word-Sense Disambiguation, *ACL WSD Workshop*, 2002, pp. 74–80.
9. Le, C. A., and Shimazu Akira, High Word Sense Disambiguation using Naive Bayesian classifier with rich features, *The 18th Pacific Asia Conference on Language, Information and Computation* (2004)pp. 105-113
10. Leacock, C., M. Chodorow, and G. Miller, Using corpus statistics and WordNet relations for Sense Identification, *Computational Linguistics* (1998) 147–165.
11. Mooney, R. J., Comparative experiments on Disambiguating Word Senses: An illustration of the role of bias in machine learning, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1996, pp. 82–91.
12. Ng, H. T., and H. B. Lee, Integrating multiple knowledge sources to Disambiguate Word Sense: An exemplar-based approach, *Proceedings of the 34th Annual Meeting of the Society for Computational Linguistics (ACL)*, 1996, pp. 40–47.
13. Pedersen, T., A simple approach to building ensembles of Naive Bayesian classifiers for Word Sense Disambiguation, *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2000, pp. 63–69.

[5] In the table, BW, M, NL, LC, and P respectively abbreviate for Bruce & Wiebe [1], Mooney [11], Ng & Lee [12], Leacock & Chodorow [10], and Pedersen [13].

14. Shafer, G., *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, 1976).
15. Smets, P. and R. Kennes, The transferable belief model, *Artificial Intelligence* **66** (1994) 191–234.
16. Wang, X. J., and Y. Matsumoto, Trajectory based word sense disambiguation, *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, August 2004, pp. 903–909.
17. Wang, H., and D. Bell, Extended k-nearest neighbours based on evidence theory, *The Computer Journal* **47** (6) (2004) 662–672.
18. Zadeh, L. A., Reviews of Books: A Methematical Theory of Evidence, *The AI Magazine* **5** (1984) 81–83.