Weighted Combination of Classifiers Based on Dempster-Shafer Theory and OWA Operators in Word Sense Disambiguation

Van-Nam Huynh[1], Cuong Anh Le[2], Akari Shimazu[2] and Yoshiteru Nakamori[1]
[1]School of Knowledge Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
{huynh, cuonganh}@jaist.ac.jp
[2] School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan

ABSTRACT

In this paper we discuss a framework for weighted combination of classifiers in which each individual classifier uses a distinct representation of objects to be classified. This framework is essentially based on Dempster-Shafer theory of evidence (Dempster, 1967; Shafer, 1976) and OWA operators (Yager, 1988). It is of interest to see that this framework not only yields many commonly used decision rules without some strong assumptions made in the work by Kittler et al. (1998), but also provides other new decision rules. As an application, we apply the proposed framework of classifier combination to the problem of word sense disambiguation (shortly, WSD). To this end, we experimentally design a set of individual classifiers, each of which corresponds to a distinct representation type of context considered in the WSD literature, and then the discussed combination strategies are tested on the datasets for four polysemous words, namely *interest*, *line*, *serve*, and *hard*. The experiment conducted for these four polysemous words shows significantly better results in comparison with previous studies on the same datasets.

Keywords: Computational linguistics, Classifier combination, Word sense disambiguation, OWA operator, Evidential reasoning.

1. INTRODUCTION

The ultimate goal of constructing classification systems is to achieve the best possible classification performance for the task at hand. This objective traditionally led to the development of different classification methods for any given pattern recognition problem. As observed in studies of pattern recognition systems, although one could choose one of learning systems available based on the analysis of an experimental assessment of these to hopefully achieve the best performance for a given pattern recognition problem, the set of patterns misclassified by them would not necessarily overlap [8]. This means that different classifiers may potentially offer complementary information about patterns to be classified. In other words, features and classifiers of different types complement one another in classification performance. This observation highly motivated the interest in combining classifiers during the recent years. The basic idea is to use all the classifiers, or their subset, for decision making of classification by combining their individual opinions to derive a consensus decision, instead of only relying on any a single decision making scheme.

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. A typical example of this scenario is a set of *k*-NN classifiers, each of which uses the same measurement vector but different classifier parameters (number of nearest neighbors *k*, or distance metrics used). 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. Further, an important issue in combining classifiers is the combination strategy used to derive a consensus decision. In [8], the authors proposed a common theoretical framework for combining classifiers which leads to many commonly used decision rules used in practice. This framework has been also applied to the problem of word sense disambiguation (WSD) in [12]. However, to derive these decision rules, this framework adopts several assumptions imposed on individual classifiers (for more details, see [8]) which, to our opinion, are difficult to be accepted and verified in the context of word sense disambiguation.

The issue of automatic disambiguation of word senses has been an interest and concern 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 [6], 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., [6] for a survey). During the last decades, many supervised machine learning algorithms have been used for this task, including Naive Bayesian (NB) model, decision trees, exemplar-based model, SVM, maximum entropy, etc. Especially, classifier combination for WSD has been received much attention recently from the community as well, e.g., [7], [5], [16], [9], [2], [3], [17]. In the spirit of categorizing into combination scenarios mentioned above, in the context of WSD, the work by Kilgarriff and Rosenxweig [7], Klein et al. [9], and Florian and Yarowsky [3] could be grouped into the first scenario. Whilst the work by Pedersen [17] can be considered as belonging to the second scenario, although the difference of representations here is only in terms of size of context windows. In this paper, we focus on weighted combination of classifiers in the second scenario with the discussion being put in the context of word sense disambiguation. Particularly, we discuss a framework for weighted combination of classifiers for WSD in which each individual classifier uses a distinct representation of objects to be classified. This framework is essentially based on Dempster-Shafer theory of evidence [18, 19] and OWA operators [21].

More particularly, we first consider various ways of using context in WSD as distinct representations of a polysemous word under consideration, and then all these representations are used jointly to identify the meaning of the target word. On the one hand, by considering each representation of the context as information inspired by a semantics or syntactical criterion for the purpose of word sense identification, we can apply OWA operators for aggregating multi-criteria to form an overall decision function considered as the fuzzy majority based voting strategy [13]. Essentially, we use OWA operators for classifier fusion in their semantic relation to linguistic quantifiers [22] so that we could provide a framework for combining classifiers, which also yields several commonly used decision rules for WSD but without some strong assumptions made in the work by Kittler et al. [8]. On the other hand, various ways of using the context could be considered as providing different information sources to identify the meaning of the target word. Moreover, 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. Then by considering the problem as that of weighted combination of evidence for decision making, we formulate a general rule of classifier combination based on Dempster-Shafer theory of evidence [18], adopting a probabilistic interpretation of weights. This interpretation of weights seems to be appropriate when defining weights in terms of the accuracy of individual classifiers. Note that the formulation of weighted classifier combination in terms of Dempster-Shafer theory also yields some interestingly classifier combination schemes.

Experimentally, we design a set of individual classifiers, each of which corresponds to a distinct representation type of context considered in the WSD literature, and then the proposed combination strategies are experimentally tested on the datasets for four polysemous words, namely *interest*, *line*, *serve*, and *hard*.

The paper is organized as follows. In section 2, we will recall basic notions from Dempster-Shafer theory of evidence and OWA operators. Section 3 devotes to the theoretical framework for combining classifiers in WSD based on these theories. Then an experimental study will be conducted in section 4. Finally, section 5 presents some concluding remarks.

2. PRELIMINARIES

In this section we briefly review basic notions of Dempster-Shafer (DS) theory of evidence and OWA operators.

2. 1. Dempster-Shafer Theory of Evidence

In Dempster-Shafer theory of evidence, a problem domain is often represented by a finite set Θ of mutually exclusive and exhaustive hypotheses, called *frame of discernment* [18]. 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:

(i) $m(\varnothing) = 0$, and

(ii) $\sum_{A \subseteq \Theta} m(A) = 1$.

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

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

and

$$Pl_m(A) = \sum_{B \cap A \neq \emptyset} m(B).$$

Two useful operations that play a central role in the manipulation of belief functions are *discounting* and *Dempster's rule of combination*. 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-α) as one's *discount rate*, which results in a new BPA m^α defined by

$$m^\alpha(A) = \alpha m(A), \text{ for any } A \subseteq \Theta$$
$$m^\alpha(\Theta) = (1-\alpha) + \alpha m(\Theta).$$

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

$$m_1 \oplus m_2(\emptyset) = 0,$$

$$m_1 \oplus m_2(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C), \quad (1)$$

where

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

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

2. 2. OWA Operators

The notion of ordered weighting average (shortly, OWA) operators was first introduced in [21] regarding the problem of aggregating multi-criteria to form an overall decision function. A mapping

$$F: [0,1]^n \rightarrow [0,1]$$

is called an OWA operator of dimension n if it is associated with a weighting vector $W=[w_1,\ldots,w_n]$, such that 1) $w_i \in [0,1]$ and 2) $\sum_i w_i = 1$, and

$$F(a_1,\ldots,a_n) = \sum_i w_i b_i$$

where b_i is the i-th largest element in the collection $a_1,.., a_n$.

OWA operators provide a type of aggregation operators which lay between the "and" and the "or" aggregation. As suggested by Yager [21], there exist at least two methods for obtaining weights w_i's. The first approach is to use some kind of learning mechanism. The second one is to give some semantics or meaning to the weights. Then, based on these semantics we can directly provide the values for the weights. In the following we use the semantics based on fuzzy linguistic quantifiers for the weights.

The fuzzy linguistic quantifiers were introduced by Zadeh in [22]. According to Zadeh, there are basically two types of quantifiers: absolute, and relative. Here we focus on the relative quantifiers typified by terms such as *most*, *at least half*, *as many as possible*. A relative quantifier Q is defined as a mapping Q: [0,1] → [0,1] verifying $Q(0)=0$, there exists $r \in [0,1]$ such that $Q(r)=1$, and Q is a non-decreasing function. For example, the membership function of relative quantifiers can be defined [4] as

$$Q(r) = \begin{cases} 0, & r < a \\ \dfrac{r-a}{b-a}, & a \le r \le b \\ 1, & r > b \end{cases} \quad (2a)$$

with parameters a, $b \in [0,1]$. Then, Yager [21] proposed to compute the weights w_i's based on the linguistic quantifier represented by Q as follows:

$$w_i = Q(\frac{i}{n}) - Q(\frac{i-1}{n}) \quad (2b)$$

for $i=1,\ldots, n$.

3. WEIGHTED COMBINATION OF CLASSIFIERS FOR WSD

Consider a pattern recognition problem where pattern $\mathbf{x}$ is to be assigned to one of the M possible classes c_1, c_2, …, c_M. Let us also assume that we have R classifiers corresponding to R distinct representations of the given pattern, denoted by $\mathbf{f}_1$, $\mathbf{f}_2,\ldots$, $\mathbf{f}_R$. Now, in order to utilize all the available information to make a decision on the classification, it is essential to consider all the representations of the pattern simultaneously and, according to the Bayesian theory [8], then the pattern $\mathbf{x}$ should be assigned to class c_j provided the a posteriori probability of that class is maximum, i.e.

$$j = \arg\max_k \; p(c_k \mid \mathbf{f}_1,...,\mathbf{f}_R) \quad (3)$$

Begin with the decision rule (3), under the conditional independence assumption of the representations used and the assumption that the posterior class probabilities computed by the respective classifiers do not deviate greatly from the prior ones, the authors in [8] developed a theoretical framework for combining classifiers which leads to many commonly used decision rules used in practice. However, the authors also conceded that these assumptions seem to be unrealistic in many situations.

Particularly, to our opinion, these assumptions are difficult to be accepted and verified in the context of WSD. In the following, we will focus on a framework for combining classifiers in WSD based on the DS theory and OWA operators. This framework also interestingly yields many commonly used decision rules for WSD but without the strong assumptions mentioned above.

3. 1. WSD with Multi-Representation of Context

Given a 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 [6]: 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 $\mathbf{f}_1$, $\mathbf{f}_2$,…, $\mathbf{f}_R$, 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 the context has its own type depending on which way context is used.

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 word **w**, is used by the i-th classifier. Furthermore, assume that each i-th classifier (expert) is associated with a weight α_i, $0 \le \alpha_i \le 1$, reflecting the relative confidence in or important of the classifier. In the following we will show that different semantic views of representations $\mathbf{f}_i$ associated with various interpretations of corresponding weights α_i lead to numerous various classifier combination schemes serving for identifying the sense of the target **w**.

3. 2. DS Theory Based Combination Scheme

Given a target word **w** in a context C and $\mathbf{S}=\{c_1,\ldots, c_M\}$ is the set of its possible senses. Using the vocabulary of DS theory, **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 $P(\bullet|\,\mathbf{f}_i)$ (i=1,…, R) on **S**,
- The weights α_i of the individual information sources (i=1,…, R) (Note that the constraint $\sum\alpha_i$=1 does not need to be imposed).

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

$$P(c_k \mid \mathbf{f}_1,...,\mathbf{f}_R) = \frac{1}{\sum \alpha_i} \sum_{i=1}^{R} \alpha_i P(c_k \mid \mathbf{f}_i) \qquad (4)$$

for k = 1,…,R. Then the target word **w** should be naturally assigned to the sense c_j according to the following decision rule

$$j = \arg\max_k P(c_k \mid \mathbf{f}_1,...,\mathbf{f}_R) \qquad (5)$$

However, by considering the problem as that of weighted combination of evidence for decision making, we now 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,…,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 $P(\bullet|\,\mathbf{f}_i)$ should be discounted at a discount rate of $(1-\alpha_i)$. This results in a BPA m_i verifying

$$m_i(c_k) = \alpha_i P(c_k \mid \mathbf{f}_i) \div p_{ik}, \text{ for } k = 1,\ldots,M$$
$$m_i(\mathbf{S}) = 1- \alpha_i \div p_{i\mathbf{S}}.$$

That is, the discount rate of $(1-\alpha_i)$ can not be distributed to anything else than **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

$$m = \bigoplus_{i=1}^{R} m_i \qquad (6)$$

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

By applying different combination operators for $\oplus$ in (6), we may have different aggregation schemes for obtaining the BPA m which models our belief for making the decision on the sense of **w**. In [11] we have examined two different combination strategies, called *discounting-and-orthogonal sum* and *discounting-and –averaging*, which correspond to applying Dempster's rule of combination and average operator for $\oplus$ respectively. Note that in this approach, after obtained the BPA m, we must also deal with the problem of how to make a decision based on it. Because m does not in

general provide a unique probability distribution on **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 **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 **S** derived from m for the purpose of decision making via the *pignistic transformation* [19]. That is, as in the two-level language of the so-called *transferable belief model* [19], 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.

Let us denote by DS1 and DS2 the *discounting-and -orthogonal sum* and *discounting-and-averaging combination strategies* respectively. It is of interest to note that the combination strategy DS2 is nothing but the weighted mixture of individual classifiers as defined in (4). Due to the limitation of space, the details of these could be referred to [11].

3. 2. OWA Operator Based Combination Scheme

Let us return to the problem of identifying the sense of a given word **w** as described above. As discussed on the role of context in the task of determining the most appropriate sense of **w**, each representation $\mathbf{f}_i$ of the context C can be also considered as providing the information inspired by a semantical or syntactical criterion for the purpose of word sense identification. Let us assume that we have R classifiers corresponding to R representations $\mathbf{f}_i$ of the context, each of which provides a soft decision for identifying the right sense of the target word **w** in the form of a posterior probability $P(c_k \mid \mathbf{f}_i)$, for i=1,…,R.

Under such a consideration, we now can define an overall decision function D, with the help of an OWA operator F of dimension R, which combines individual opinions to derive a consensus decision as follows:

$$D(c_k) = F_W(P(c_k \mid \mathbf{f}_1),...P(c_k \mid \mathbf{f}_R)) = \sum_{i=1}^{R} w_i p_i \qquad (7)$$

where p_i is the i-th largest element in the collection $P(c_k|\mathbf{f}_1),\ldots,P(c_k|\mathbf{f}_R)$, and W=[$w_1,\ldots,w_R$] is a weighting vector semantically associated with a fuzzy linguistic quantifier. Then, the fuzzy majority based voting strategy suggests that the word **w** should be assigned to class c_j provided that $D(c_j)$ is maximum, namely

$$j = \arg\max_k D(c_k) \qquad (8)$$

It should be worth mentioning that the use of OWA operators in classifier combination has been studied, for example, in [10]. In this work we use OWA operators for classifier fusion in their semantic relation to linguistic quantifiers so that we could provide a framework for combining classifiers, which also yields several commonly used decision rules but without some strong assumptions made in the work by Kittler et [8].

As studied in [21], using Zadeh's concept of linguistic quantifiers [22] and Yager's idea of associating their semantics to various weighting vectors W, we can obtain many commonly used decision rules as following.

Max Rule. First let us use the quantifier *there exists* which can be relatively represented as a fuzzy set Q of [0,1] such that $Q(r) = 0$, for $r < 1/R$ and $Q(r)=1$, for $r \geq 1/R$. We then obtain from (2b) the weighting vector W=[1,0,…,0], which yields from (7) and (8) the Max Decision Rule as

$$j = \arg\max_k \max_i P(c_k \mid \mathbf{f}_i)$$

Min Rule. Similarly, if we use the quantifier *for all* which can be relatively represented as a fuzzy set Q of [0,1] such that $Q(1) = 1$, and $Q(r)=0$, for $r \neq 1$ [25]. We then obtain from (2b) the weighting vector W=[0,…,0,1], which yields from (7) and (8) the Min Decision Rule as

$$j = \arg\max_k \min_i P(c_k \mid \mathbf{f}_i)$$

Median Rule. In order to have the Median decision rule, we use the absolute quantifier *at least one* which can be equivalently represented as a relative quantifier with the parameter pair (0,1) for the membership function Q in (2a). Then we obtain from (2b) the weighting vector W=[1/R,…,1/R], which from (7) and (8) leads to the median decision rule as

$$j = \arg\max_k [\tfrac{1}{R} \sum_{i=1}^{R} P(c_k \mid \mathbf{f}_i)]$$

Fuzzy Majority Voting Rules. We now use the relative quantifier *at least half*} with the parameter pair (0,0.5) for the membership function Q in (2a). Then, depending on a particular value of R, we can obtain from (2b) the corresponding weighting vector W=[$w_1,\ldots,w_R$] for the decision rule, denoted by FM1, as:

$$j = \arg\max_k [\sum_{i=1}^{R} w_i p_i]$$

where p_i is the i-th largest element in the collection $P(c_k|\mathbf{f}_1),\ldots,P(c_k|\mathbf{f}_R)$.

Similarly, we can also use the relative quantifier *as many as possible* with the parameter pair (0.5,1) for the membership function Q in (2a) to obtain the

corresponding decision rule, denoted by FM2.

Interestingly also, from the following relation

$$\prod_{i=1}^{R} P(c_k \mid \mathbf{f}_i) \le \min_i P(c_k \mid \mathbf{f}_i) \le \sum_i w_i p_i$$
$$\le \max_i P(c_k \mid \mathbf{f}_i) \le \sum_{i=1}^{R} P(c_k \mid \mathbf{f}_i)$$

it suggests that the Max and Min decision rules can be approximated by the upper or lower bounds appropriately. Especially, under the assumption of equal priors, the decision rule derived from (3) (see [8]) simplifies to the Product rule, which is a lower approximation of the Min rule, while approximating Max rule by the upper bound yields the Sum rule.

In addition, from the classical voting strategy, we can also obtain the following decision rule.

Majority Vote Rule. Majority voting follows a simple rule as: it will vote for the class which is chosen by maximum number of individual classifiers. This can be done by hardening the a posteriori probabilities $P(c_k \mid \mathbf{f}_i)$ in terms of functions Δ_{ki} defined as follows:

$$\Delta_{ki} = \begin{cases} 1, & \textbf{if } P(c_k|f_i) = \max_j P(c_j \mid \mathbf{f}_i) \\ 0, & \textbf{otherwise} \end{cases}$$

then the right class (sense) c_j is determined as follows:

$$j = \arg\max_k \sum_i \Delta_{ki}$$

4. AN EXPERIMENTAL STUDY

In this section we will design an experiment to test the classifier combination schemes discussed.

4. 1. Representations of Context for WSD

As mentioned above, context plays an essentially important role in WSD and 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. Ng and Lee [16] proposed the use of more linguistic knowledge resources including topic context, collocation of words, and a syntactic relationship verb-object, which then became popular resources for determining word sense in many papers. In [14], 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 strategies, 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 [17] and Wang and Matsumoto [20].

On the other hand, we observe that two of the most important information sources for determining the sense of a polysemous word are the topic of context and relational information representing the structural relations between the target word and the surrounding words in a local context. Under such an observation, we have experimentally designed five kinds of representation defined as follows: $\mathbf{f}_1$ is a set of unordered words in the large context; $\mathbf{f}_2$ is a set of words assigned with their positions in the local context; $\mathbf{f}_3$ is a set of part-of-speech tags assigned with their positions in the local context; $\mathbf{f}_4$ is a set of collocations of words; $\mathbf{f}_5$ is a set of collocations of part-of-speech tags. Symbolically, we have

- $\mathbf{f}_1 = \{w_{-n_1}, ..., w_{-2}, w_{-1}, w_1, w_2, ..., w_{n_1}\}$
- $\mathbf{f}_2 = \{(w_{-n_2}, -n_2), ..., (w_{-1}, -1), (w_1, 1), ..., (w_{n_2}, n_2)\}$
- $\mathbf{f}_3 = \{(p_{-n_3}, -n_3), ..., (p_{-1}, -1), (p_1, 1), ..., (p_{n_3}, n_3)\}$
- $\mathbf{f}_4 = \{w_{-l} ... w_{-1} \mathbf{w} w_1 ... w_r \mid l + r \le n_4\}$
- $\mathbf{f}_5 = \{p_{-l} ... p_{-1} \mathbf{w} p_1 ... p_r \mid l + r \le n_5\}$

where w_i is the word at position i in the context of the ambiguous word $\mathbf{w}$ and p_i be the part-of-speech tag of w_i, with the convention that the target word $\mathbf{w}$ appears precisely at position 0 and i will be negative (positive) if w_i appears on the left (right) of $\mathbf{w}$. 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 n_i of local context as 3 for remaining representations.

4. 2. 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 [17], Ng and Lee [16], Bruce and Wiebe [1], and Leacock, Chodorow and Miller [14]. We have obtained those data from Pedersen's homepage[1]. 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.

4. 3. Experimental Results

In the experiment, we obtain the results that are the average of 5 results from 10-folds cross validation. Data

[1] http://www.d.umn.edu/~tpederse/data.html

included four datasets corresponding to four polysemous words *interest*, *line*, *hard*, and *serve*, were tested based on multi-representation of context as defined in the preceding section.

Table 1 shows the experimental results obtained by using various strategies of classifier combination developed in Section 3 and the best results obtained by individual classifiers respectively. It is of interest to note that Majority Voting, which is widely used in many studies of combining classifiers, may not be a good choice for classifier combination in WSD.

Table 2 shows the comparison of results from the best classifier combination with previous WSD studies, which were also tested on the four words. It is shown that the best classifier combination based on multi-representation of context gives the highest accuracy on all the four words, except for the word *line* where Pedersen's method does better.

5. CONCLUSIONS

In this paper we have discussed and formalized various ways of using context in WSD as distinct representations of a polysemous word under consideration, and then all these representations are used jointly to identify the meaning of the target word. This consideration allowed us to develop a framework for combining classifiers based on DS theory and the notion of OWA operators with the help of fuzzy majorities. Interestingly, this framework also yields many commonly used decision rules for WSD, without assumptions imposed on individual classifiers as done in [10]. We also experimentally explored all developed combination strategies on the datasets for four polysemous words, namely *interest*, *line*, *serve*, and *hard*, which are used in numerous comparative studies of word sense disambiguation methodologies. It has been also shown that multi-representation of context could significantly improve 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.

However, more reliable results would be needed to strongly support the claim of improvement in WSD by the developed framework. We are planning to revise and test the classifier combination schemes discussed in this paper with the Senseval data [7] for that purpose.

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]. Escudero, G., L. Marquez, and G. Rigau, Boosting applied to word sense disambiguation, *Proceedings of the 11th European Conference on Machine Learning* (ECML), 2000, pp. 129-141.

[3]. Florian, R., and D. Yarowsky, Modeling consensus: Classifier combination for word sense disambiguation, *Proceedings of EMNLP* 2002, pp. 25-32.

[4]. Herrera, F. and J.L. Verdegay, A linguistic decision process in group decision making, *Group Decision Negotiation* **5** (1996) 165-176.

[5]. 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.

[6]. Ide, N., J. Veronis, Introduction to the special issue on word sense disambiguation: The state of the art, *Computational Linguistics* **24** (1998) 1-40.

[7]. Kilgarriff, A., and J. Rosenzweig, Framework and results for English SENSEVAL, *Computers and the Humanities* **36** (2000) 15-48.

[8]. 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.

[9]. 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.

[10]. Kuncheva, L.I., Combining classifiers: Soft computing solutions, in: S.K. Pal, A. Pal, *Pattern Recognition: From Classical to Modern Approaches*, World Scientific, 2001, pp. 427-451.

[11]. Le, C.A., V.N. Huynh, A. Shimazu, An evidential reasoning approach to weighted combination of classifiers for word sense disambiguation, *MLDM 2005*, P. Perner and A. Imiya (Eds.), Springer-Verlag, LNCS **3587**, pp. 516-525.

[12]. Le, C.A., V.N. Huynh, A. Shimazu, Combining classifiers with multi-representation of context in word sense disambiguation, *PAKDD 2005*, T.B. Ho et al. (Eds.), Springer-Verlag, LNAI **3518**, pp. 262-268.

[13]. Le, C.A., V.N. Huynh, H.C. Dam, A. Shimazu, Combining classifiers based on OWA operators with an application to word sense disambiguation,

RSFDGrC 2005, D. Slezak et al. (Eds.), Springer-Verlag, LNAI **3641**, pp. 512-521.

[14]. Leacock, C., M. Chodorow, and G. Miller, Using corpus statistics and WordNet relations for Sense Identification, *Computational Linguistics* **24** (1998) 147-165.

[15]. 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.

[16]. 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.

[17]. Pedersen, T., A simple approach to building ensembles of Naïve Bayesian classifiers for word sense disambiguation, *Proceedings of the North American Chapter of the Association for Computational Linguistics* (NAACL), 2000, pp. 63-69.

[18]. Shafer, G., *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, 1976).

[19]. Smets, P. and R. Kennes, The transferable belief model, *Artificial Intelligence* **66** (1994) 191-234.

[20]. 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.

[21]. Yager, R. R., On ordered weighted averaging aggregation operators in multicriteria decision making, *IEEE Transactions on Systems, Man, and Cybernetics* **18** (1988) 183-190.

[22]. Zadeh, L.A., A computational approach to fuzzy quantifiers in natural languages, *Computers and Mathematics with Applications* **9** (1983) 149-184.

Table 1. Experimental Results

	Best individual classifier (%)	Max (%)	Min (%)	Median (%)	Majority (%)	DS1 (%)	DS2 (%)	FM1 (%)	FM2 (%)
interest	87.0	89.6	89.9	90.5	88.7	**91.2**	90.7	90.2	89.7
line	82.8	86.9	**87.2**	84	79.8	**87.2**	85.6	84.3	82.7
hard	90.2	89.8	89.2	91	90.4	**91.6**	91.3	91	90.9
serve	84.4	87.7	88.1	88.6	85.4	**89.7**	89.1	89	88.8

Table 2. The comparison with previous studies

	Bruce & Wiebe [1] (%)	Mooney [15] (%)	Ng & Lee [16] (%)	Leacock, Chodorow & Miller [14] (%)	Pedersen [17] (%)	Best combiner (%)
interest	78	-	87	**-**	89	**91.2**
line	-	72	-	84	**88**	87.2
hard	-	-	-	83	-	**91.6**
serve	-	-	-	83	-	**89.7**