Combining Classifiers for Word Sense Disambiguation Based on Dempster-Shafer Theory and OWA Operators

Cuong Anh Le, Van-Nam Huynh, Akira Shimazu, Yoshiteru Nakamori

Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
Email: huynh@jaist.ac.jp (V.-N. Huynh)

Abstract

In this paper, we discuss a framework for weighted combination of classifiers for word sense disambiguation (WSD). This framework is essentially based on Dempster-Shafer theory of evidence (Shafer, 1976) and ordered weighted averaging (OWA) operators (Yager, 1988). We first determine various kinds of features which could provide complementarily linguistic information for the context, and then combine these sources of information based on Dempster's rule of combination and OWA operators for identifying the meaning of a polysemous word. 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 and compared on English lexical samples of Senseval-2 and Senseval-3.

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

1 Introduction

The issue of automatic disambiguation of word senses has been an interest and concern since the 1950s, and is still one of the most important open problems in natural language processing (NLP) [24]. 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 [11], this is an "intermediate task" necessarily to accomplish most NLP tasks such as grammatical analysis and lexicography in linguistic studies, or

machine translation, man-machine communication, message understanding in language understanding applications. Besides these applications, WSD may also have potential uses in other applications involving data and knowledge engineering such as information retrieval, information extraction and text mining [1]. More particularly, in information retrieval (IR), the ambiguity has to be resolved in some queries; for example, given the query "*depression*", should the system return documents about illness, weather systems, or economics? Though current IR systems do not use explicitly WSD and rely on the user typing enough context in the query in order to only retrieve documents relevant to the intended sense (e.g. "*tropical depression*"), early experiments suggested that reliable IR would require at least 90% disambiguation accuracy for explicit WSD to be of benefit [28]. In addition, WSD has been more recently shown to improve cross-lingual IR and document classification [3,4,31]. On the other hand, in information extraction and text mining, WSD is required for the accurate analysis of text in many applications. For instance, an intelligence gathering system might require the flagging of all the references illegal *drugs*, rather than medical *drugs*. More generally, the Semantic Web requires automatic annotation of documents according to a reference ontology: all textual references must be resolved to the right concepts and event structures in the ontology (see [5]). Named-entity classification, co-reference determination, and acronym expansion can also be cast as WSD problems for proper names. WSD is only beginning to be applied in these areas.

Since its inception, many approaches have been proposed for WSD in the literature (see [11] for a survey). During the last two decades, many supervised machine learning algorithms have been used for this task, including Naive Bayesian (NB) model, decision trees, exemplar-based model, support vector machines, maximum entropy models, etc. On the other hand, 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 the pattern recognition problem at hand, the set of patterns misclassified by them would not necessarily overlap [14]. 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, with particularly application to WSD as in [6,7,10,12,15,27,32].

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, while in the second scenario, each classifier uses its own representation of the input pattern. An important application of combining classifiers in the second scenario is the possibility to integrate physically different types of features. In addition, an important issue in combining classifiers is what

combination strategy should be used to derive a consensus decision. In [14], 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 WSD in [17,20]. In this paper [1], we focus on the combination of classifiers according to the second scenario with the discussion being put in the context of WSD. Particularly, 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 based on Dempster-Shafer (DS) theory of evidence [29] and OWA operators [33].

In [2], Al-Ani and Deriche have proposed a new technique for combining classifiers using DS theory, in which different classifiers correspond to different feature sets. In their approach, the distance between the output classification vector provided by each single classifier and a reference vector is used to estimate basic probability assignments (BPAs). These BPAs are then combined making use of Dempster's rule of combination to obtain a new output vector that represents the combined confidence in each class label. Different from their approach, we directly use the output classification vectors of individual classifiers to define the corresponding BPAs, making use of the discount operation in DS theory and then combine the resulted BPAs to obtain the final BPA for making the decision of classification. 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, 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 DS theory of evidence [29], 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.

On the other 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 multicriteria to form an overall decision function considered as the fuzzy majority based voting strategy. It should be worth mentioning that the use of OWA operators in classifier combination has been studied, for example, in [16]. In this paper, however, we use OWA operators for classifier fusion in their seman-

[1] This paper is a revised, unified and substantially expanded version of the papers presented at MLDM'2005 [18] and RSFDGrC'2005 [19].

tic 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 al. [14].

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 English lexical samples of Senseval-2 and Senseval-3. The rest of this 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 DS theory of evidence and OWA operators.

2.1 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* [29]. 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* [29]. The discounting operation is used when a source of information provides a BPA m, but one knows that this source has probability α of reliability. 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 $\kappa < 1$.

2.2 OWA Operators

The notion of OWA operators was first introduced in [33] regarding the problem of aggregating multi-criteria to form an overall decision function. A mapping

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

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

$$F(a_1, \ldots, a_n) = \sum_{i=1}^{n} w_i b_i$$

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

OWA operators provide a type of aggregation operators which lay between the "and" and the "or" aggregation. As suggested by Yager [33], there exist at least two methods for obtaining weights w_i's. The first approach is to use some kind of learning mechanism. That is, we use some sample data, arguments and associated aggregated values and try to fit the weights to this collection of sample data. The second approach 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 [36]. 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 $Q : [0, 1] \to [0, 1]$ 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 [9] as

$$Q(r) = \begin{cases} 0 & \text{if } r < a \\ \frac{r-a}{b-a} & \text{if } a \leq r \leq b \\ 1 & \text{if } r > b \end{cases} \tag{5}$$

with parameters $a, b \in [0, 1]$.

Then, Yager [33] proposed to compute the weights w_i's based on the linguistic quantifier represented by Q as follows:

$$w_i = Q\left(\frac{i}{n}\right) - Q\left(\frac{i-1}{n}\right), \text{ for } i = 1, \ldots, n. \tag{6}$$

3 Weighted Combination Of Classifiers For WSD

Consider a pattern recognition problem where pattern $\mathbf{w}$ is to be assigned to one of the M possible classes $c_1, c_2, \ldots, c_M$. 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 [14], then the pattern $\mathbf{w}$ should be assigned to class c_j provided the a posteriori probability of that class is maximum, i.e.

$$j = \arg\max_k P(c_k|\mathbf{f}_1, \ldots, \mathbf{f}_R) \tag{7}$$

Begin with the decision rule (7), 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 [14] developed a theoretical framework for combining classifiers which leads to many commonly used decision rules used in practice. At the same time, 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 $\mathbf{w}$, 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 [11]: 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 $\mathbf{w}$, 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 $\mathbf{w}$. Clearly, each $\mathbf{f}_i$ can be also considered as a semantical representation of $\mathbf{w}$. Each representation $\mathbf{f}_i$ of 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 $\mathbf{w}$, is used by the i-th classifier (see Fig. 1). 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 dif-

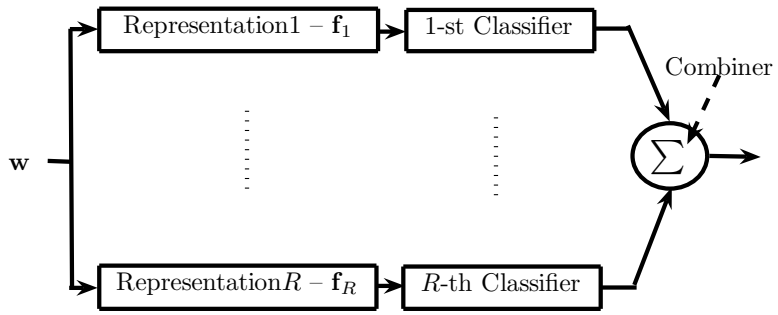

Fig. 1. A Scheme of Multi-Classifier Combination

ferent 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 or importance 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 classifier combination schemes serving for identifying the sense of the target $\mathbf{w}$.

3.2 DS Theory Based Combination Scheme

Given a target word $\mathbf{w}$ 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 $\mathbf{w}$ 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 $(i = 1, \ldots, R)$ [2].

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

$$P(c_k) = \frac{1}{\sum_i \alpha_i} \sum_{i=1}^{R} \alpha_i P(c_k|\mathbf{f}_i), \text{ for } k = 1, \ldots, M \tag{8}$$

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

Then the target word $\mathbf{w}$ should be naturally assigned to the sense c_j according to the following decision rule

$$j = \arg\max_k P(c_k) \tag{9}$$

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 $(i = 1, \ldots, R)$ 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 defined by

$$m_i(\{c_k\}) = \alpha_i P(c_k|\mathbf{f}_i) \triangleq p_{i,k}, \text{ for } k = 1, \ldots, M \tag{10}$$

$$m_i(\mathcal{S}) = 1 - \alpha_i \triangleq p_{i,\mathcal{S}} \tag{11}$$

$$m_i(A) = 0, \forall A \in 2^{\mathcal{S}} \setminus \{\mathcal{S}, \{c_1\}, \ldots, \{c_M\}\} \tag{12}$$

That is, the discount rate of $(1 - \alpha_i)$ should 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

$$m = \bigoplus_{i=1}^{R} m_i \tag{13}$$

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

By applying different combination operations for $\oplus$, we may have different aggregation schemes for obtaining the BPA m which models our belief for making the decision on the sense of $\mathbf{w}$. 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* as stated in [30], we may define a probability function P_m on $\mathcal{S}$ derived from m for the purpose of decision making via the so-called *pignistic transformation*. That is, as in the two-level language of the so-called *transferable belief model* [30], the aggregated BPA m itself representing the belief

is obtained 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.

3.2.1 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 [29], 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 (13) 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 [34]. 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

$$p_{I(i),k} \triangleq m_{I(i)}(\{c_k\}), \text{ for } k = 1, \ldots, M \tag{14}$$

$$p_{I(i),\mathcal{S}} \triangleq m_{I(i)}(\mathcal{S}) \tag{15}$$

With these notations and (10)–(11), 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

$$p_{I(i+1),k} = \frac{1}{\kappa_{I(i+1)}}[p_{I(i),k}p_{i+1,k} + p_{I(i),k}p_{i+1,\mathcal{S}} + p_{I(i),\mathcal{S}}p_{i+1,k}] \tag{16}$$

$$p_{I(i+1),\mathcal{S}} = \frac{1}{\kappa_{I(i+1)}}(p_{I(i),\mathcal{S}}p_{i+1,\mathcal{S}}) \tag{17}$$

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

by

$$\kappa_{I(i+1)} = \left[1 - \sum_{j=1}^{M} \sum_{\substack{k=1 \\ k \neq j}}^{M} p_{I(i),j} p_{i+1,k} \right] \tag{18}$$

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

$$P_m(c_k) = m(\{c_k\}) + \frac{1}{M} m(\mathcal{S}) \text{ for } k = 1, \ldots, M \tag{19}$$

and we have the following decision rule:

$$j = \arg\max_k P_m(c_k) \tag{20}$$

It would be interesting to note that an issue may arise with the orthogonal sum operation, and is in using 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 [37]. 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.

3.2.2 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

$$m(A) = \frac{1}{R} \sum_{i=1}^{R} m_i(A) \tag{21}$$

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

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

$$m(\mathcal{S}) = 1 - \frac{\sum_{i=1}^{R} \alpha_i}{R} \triangleq 1 - \overline{\alpha} \tag{23}$$

$$m(A) = 0, \forall A \in 2^{\mathcal{S}} \setminus \{\mathcal{S}, \{c_1\}, \ldots, \{c_M\}\} \tag{24}$$

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

$$m(\{c_k\}) = \frac{1}{\sum_i \alpha_i} \sum_{i=1}^{R} \alpha_i P(c_k|\mathbf{f}_i), \text{ for } k = 1, \ldots, M \tag{25}$$

$$m(A) = 0, \forall A \in 2^{\mathcal{S}} \setminus \{\{c_1\}, \ldots, \{c_M\}\} \tag{26}$$

which interestingly turns out to be the weighted mixture of individual classifiers as defined in (8). Then we have the decision rule (9).

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 defined by (22)–(24) via the pignistic transformation.

3.3 OWA Operator Based Combination Scheme

Let us return to the problem of identifying the sense of a given word $\mathbf{w}$ as described above. As discussed on the role of context in the task of determining the most appropriate sense of $\mathbf{w}$, 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 $\mathbf{w}$ in the form of a posterior probability $P(c_k|\mathbf{f}_i)$, for $i = 1, \ldots, R$.

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:

$$D(c_k) = F(P(c_k|\mathbf{f}_1), \ldots, P(c_k|\mathbf{f}_R)) = \sum_{i=1}^{R} w_i p_i \tag{27}$$

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 target word $\mathbf{w}$ should be assigned to class c_j provided that $D(c_j)$ is maximum, namely

$$j = \arg\max_k D(c_k) \tag{28}$$

As studied in [33], using Zadeh's concept of linguistic quantifiers and Yager's idea of associating their semantics to various weighting vectors W, we can obtain many commonly used decision rules as following.

3.3.1 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 (6) the weighting vector $W = [1, 0, \ldots, 0]$, which yields from (27) and (28) the Max Decision Rule as

$$j = \arg\max_k \left[\max_i P(c_k|\mathbf{f}_i)\right] \tag{29}$$

3.3.2 Min Rule.

Similarly, if we use the quantifier *for all* which can be defined as a fuzzy set Q of $[0, 1]$ such that $Q(1) = 1$ and $Q(r) = 0$, for $r \neq 1$ [33]. We then obtain from (6) the weighting vector $W = [0, \ldots, 0, 1]$, which yields from (27) and (28) the Min Decision Rule as

$$j = \arg\max_k \left[\min_i P(c_k|\mathbf{f}_i)\right] \tag{30}$$

3.3.3 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 (5). Then we obtain from (6) the weighting vector $W = [1/R, \ldots, 1/R]$, which from (27) and (28) leads to the median decision rule as:

$$j = \arg\max_k \left[\frac{1}{R}\sum_{i=1}^{R} P(c_k|\mathbf{f}_i)\right] \tag{31}$$

3.3.4 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 (5) as graphically depicted in Fig. 2. Then, depending on a particular value of R, we can obtain from (6) the corresponding weighting vector $W = [w_1, \ldots, w_R]$ for the decision rule, denoted by FM1, as:

$$j = \arg\max_k \left[\sum_{i=1}^{R} w_i p_i\right] \tag{32}$$

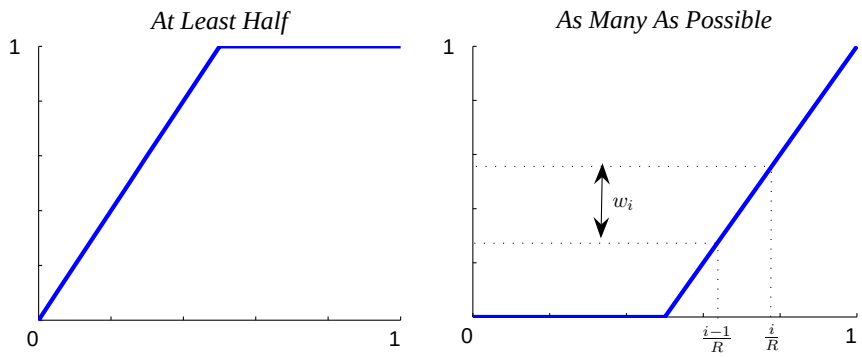

Fig. 2. Linguistic Quantifiers

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 (5) (graphically, see Fig. 2) to obtain the corresponding decision rule, denoted by FM2.

Interestingly also, from the following relation

$$\prod_{i=1}^{R} P(c_k|\mathbf{f}_i) \leq \min_{i=1}^{R} P(c_k|\mathbf{f}_i) \leq \sum_{i=1}^{R} w_i p_i \leq \max_{i=1}^{R} P(c_k|\mathbf{f}_i) \leq \sum_{i=1}^{R} P(c_k|\mathbf{f}_i) \tag{33}$$

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 (7) (see [14]) 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.

3.3.5 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 $P(c_k|\mathbf{f}_i)$ in terms of functions Δ_{ki} defined as follows:

$$\Delta_{ki} = \begin{cases} 1, & \text{if } P(c_k|\mathbf{f}_i) = \max\limits_{j} P(c_j|\mathbf{f}_i) \\ 0, & \text{otherwise} \end{cases}$$

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

$$j = \arg\max_{k} \sum_{i} \Delta_{ki} \tag{34}$$

4 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 previously, context representation plays an essentially important role in WSD. For predicting senses of a word, information usually used in all studies is the topic context which is represented by bag of words. In [26], Ng and Lee proposed a use of more linguistic knowledge resources that then became popular 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 [21], 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. In the second scenario of classifier combination, topical context with different sizes of context windows is usually used for creating different representations of a polysemous word, such as in Pedersen [27] and Wang and Matsumoto [32].

Particularly, Pedersen [27] considered several context windows on both the left and the right and grouped them into three kinds: small with window sizes 0, 1, 2; medium with window sizes 3, 4, 5; and large with window sizes 10, 25, 50. There were 81 different representations generated from combining between left and right window sizes. Finally, the best of each kind according to the majority voting procedure is then chosen. Wang and Matsumoto [32] also used only the content words in various window sizes with different left and right window sizes being (1, 2, 3, 4, 5, 6, 10, 15, 20). In this paper, for the comparison with our own representation of context, we also carry out an experiment on Pedersen's representation of context. We borrowed this feature space division from Pedersen [27] and used the maximum window size in each kind, consequently nine different representations were generated based on nine different combinations of left and right windows as follows: (2, 2), (2, 5), (2, 50), (5, 2), (5, 5), (5, 50), (50, 2), (50, 5), and (50, 50).

For context representation, 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 four kinds of representation with six feature sets defined as follows: $\mathbf{f}_1$ is a set of collocations of words; $\mathbf{f}_2$ 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; $\mathbf{f}_4, \mathbf{f}_5$ and $\mathbf{f}_6$ are sets of unordered words in the large context with different windows: small, median and large respectively. Symbolically, we have

$$\mathbf{f}_1 = \{\mathbf{w}_{-l} \cdots \mathbf{w}_{-1}\mathbf{w}\mathbf{w}_1 \cdots \mathbf{w}_r | l + r \leq n_1\}$$
$$\mathbf{f}_2 = \{(\mathbf{w}_{-n_2}, -n_2), \ldots, (\mathbf{w}_{-1}, -1), (\mathbf{w}_1, 1), \ldots, (\mathbf{w}_{n_2}, n_2)\}$$
$$\mathbf{f}_3 = \{(p_{-n_3}, -n_3), \ldots, (p_{-1}, -1), (p_1, 1), \ldots, (p_{n_3}, n_3)\}$$
$$\mathbf{f}_i = \{\mathbf{w}_{-n_i}, \ldots, \mathbf{w}_{-2}, \mathbf{w}_{-1}, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{n_i}\} \; for \; i = 4, 5, 6$$

where $\mathbf{w}_i$ 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 $\mathbf{w}_i$, with the convention that the target word $\mathbf{w}$ appears precisely at position 0 and i will be negative (positive) if $\mathbf{w}_i$ appears on the left (right) of $\mathbf{w}$.

In the experiment, we set $n_1 = 3$ (maximum of collocations), $n_2 = 5$ and $n_3 = 5$ (windows size for local context). For topic context, we foresee three different window sizes: $n_4 = 5$ (small), $n_5 = 10$ (median), and $n_6 = 100$ (large). Topical context is represented by a set of content words that includes nouns, verbs and adjectives in a certain window size. Note that after these words being extracted, they will be converted into their root morphology forms for use. 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 the unordered words in a local context may also contain structure information, but collocations and words as well as part-of-speech tags assigned with their positions may bring richer information.

4.2 Test Data

Concerning evaluation exercises in automatic WSD, three corpora so-called Senseval-1, Senseval-2 and Senseval-3 have been built on the occasion of three corresponding workshops held in 1998, 2001, and 2004 respectively. There are different tasks in these workshops with respect to different languages and/or the objectives of disambiguating single-word or all-words in the input. In this paper, the investigated combination rules will be tested on English lexical samples of Senseval-2 and Senseval-3. These two datasets are more precise than the one in Senseval-1 and widely used in current WSD studies.

A total of 73 nouns, adjectives, and verbs are chosen in Senseval-2 with the sense inventory is taken from WordNet 1.7. The data came primarily from the Penn Treebank II corpus, but was supplemented with data from the British National Corpus whenever there was an insufficient number of Treebank instances (see [13] for more detail). Examples in English lexical sample

of Senseval-3 are extracted from the British National Corpus. The sense inventory used for nouns and adjectives is taken from WordNet 1.7.1, which is consistent with the annotations done for the same task during Senseval-2. Verbs are instead annotated with senses from Wordsmyth [3]. There are 57 nouns, adjectives, and verbs in this data (see [23] for more detail).

In these datasets, each polysemous word is associated with its corresponding training dataset and test dataset. The training dataset contains sense-tagged examples, i.e. in each example the polysemous word is assigned with the right sense. The test dataset contains sense-untagged examples, and the evaluation is based on a key-file, i.e. the right senses of these test examples are listed in this file. The evaluation used here follows the proposal in [22], which provides a scoring method for exact matches to fine-grained senses as well as one for partial matches at a more coarse-grained level. Note that, like most related studies, we just compute the fine-grained score in the following experiments.

4.3 Experimental Results

Table 1 shows the experimental results conducted on Senseval-3 which are obtained by using various strategies of classifier combination developed in Section 3 and the results obtained by individual classifiers respectively. In the table, C_i $(i = 1, \ldots, 6)$ respectively represent six individual classifiers corresponding to the six feature sets $\mathbf{f}_i$ $(i = 1, \ldots, 6)$. The columns denoted by "Max", "Min", "Med", "FM1", and "FM2" show the results obtained by applying the max, min, median, FM1, and FM2 rules, respectively. Further, DS_1 denotes the Dempster rule of combination with discounting factor, while DS_2 stands for the Dempster rule of combination without discounting factor, or equivalently, $\alpha_i = 1$ for $i = 1, \ldots, 6$. In order to estimate reliable probability α_i for classifier C_i $(i = 1, \ldots, 6)$ used in DS_1 rule, we implement a 10-fold cross validation on the training data and set the obtained accuracy as α_i. The obtained results through the words in Senseval-2 show that in most cases combining classifiers gives better results in comparison with individual classifiers. On average, the best combination rule gives much better result than the one obtained from the best individual classifier (72.4% of DS_1 in comparison with 64.1% of C_4).

Table 2 shows an experimental comparison between the combination strategies discussed above and the others studied in the literature including majority voting, weighted voting, a stacking method using a maximum entropy model and Naive Bayesian combination rule. These first three methods were used for WSD in [15], and the Naive Bayesian combination rule was presented in [14]. At the same time, to see how the representation of context affects

[3] http://www.wordsmyth.net/

Table 1
Experimental Results on Senseval-3 Data

	C_1 (%)	C_2 (%)	C_3 (%)	C_4 (%)	C_5 (%)	C_6 (%)	DS_1 (%)	DS_2 (%)	Max (%)	Min (%)	Med (%)	FM1 (%)	FM2 (%)
activate.v	71.1	73.7	65.8	86.0	75.4	71.9	**87.7**	79.8	76.3	**87.7**	83.3	82.5	85.1
add.v	81.8	80.3	73.5	61.4	59.8	68.2	80.3	**87.9**	81.1	75.0	84.1	84.8	80.3
appear.v	70.7	66.9	66.2	59.4	57.1	58.6	69.9	72.9	72.9	68.4	69.2	69.2	**75.2**
argument.n	46.8	47.7	48.6	46.8	48.6	42.3	46.8	**53.2**	47.7	51.4	44.1	43.2	51.4
arm.n	83.5	85.7	84.2	88.0	85.7	88.0	**91.7**	91.0	90.2	91.0	89.5	89.5	90.2
ask.v	58.0	**64.1**	59.5	28.2	37.4	34.4	54.2	61.8	61.1	46.6	61.8	61.8	55.0
atmosphere.n	55.6	51.9	56.8	70.4	67.9	66.7	**75.3**	64.2	55.6	72.8	64.2	65.4	74.1
audience.n	67.0	81.0	77.0	72.0	82.0	76.0	83.0	84.0	71.0	77.0	86.0	85.0	**88.0**
bank.n	65.2	70.5	59.8	83.3	75.8	77.3	**84.8**	78.0	77.3	84.1	82.6	81.8	81.8
begin.v	60.8	53.2	55.7	46.8	55.7	58.2	63.3	62.0	60.8	60.8	62.0	62.0	**64.6**
climb.v	55.2	62.7	59.7	64.2	62.7	67.2	77.6	77.6	67.2	73.1	**80.6**	79.1	77.6
decide.v	**77.4**	72.6	69.4	56.5	66.1	66.1	74.2	**77.4**	75.8	69.4	75.8	**77.4**	74.2
degree.n	69.5	72.7	66.4	66.4	78.1	71.9	**85.2**	78.9	71.1	81.3	82.0	80.5	83.6
difference.n	56.1	50.0	41.2	50.9	47.4	41.2	**62.3**	58.8	55.3	58.8	58.8	57.9	61.4
different.a	48.0	48.0	40.0	38.0	38.0	34.0	46.0	54.0	52.0	44.0	**58.0**	54.0	50.0
difficulty.n	34.8	34.8	39.1	39.1	30.4	30.4	**47.8**	43.5	43.5	30.4	**47.8**	43.5	39.1
disc.n	36.0	47.0	36.0	79.0	60.0	65.0	**84.0**	53.0	46.0	81.0	67.0	62.0	71.0
eat.v	82.8	77.0	77.0	86.2	83.9	81.6	86.2	**90.8**	85.1	86.2	**90.8**	**90.8**	88.5
encounter.v	60.0	63.1	50.8	70.8	**75.4**	67.7	72.3	70.8	60.0	72.3	70.8	70.8	70.8
expect.v	**85.9**	75.6	74.4	67.9	71.8	70.5	80.8	83.3	84.6	76.9	82.1	**85.9**	80.8
express.v	52.7	50.9	60.0	41.8	**70.9**	67.3	52.7	58.2	50.9	49.1	61.8	60.0	63.6
hear.v	46.9	62.5	53.1	59.4	50.0	59.4	**68.8**	59.4	56.3	**68.8**	62.5	56.3	65.6
hot.a	74.4	65.1	69.8	76.7	79.1	**81.4**	**81.4**	**81.4**	79.1	79.1	**81.4**	**81.4**	**81.4**
image.n	51.4	51.4	37.8	71.6	56.8	58.1	70.3	59.5	55.4	**74.3**	64.9	64.9	63.5
important.a	36.8	31.6	**42.1**	31.6	26.3	36.8	36.8	31.6	36.8	31.6	31.6	36.8	36.8
interest.n	67.7	64.5	54.8	64.5	63.4	64.5	79.6	78.5	77.4	76.3	80.6	79.6	**81.7**
judgment.n	43.8	46.9	40.6	56.3	40.6	43.8	56.3	43.8	46.9	**59.4**	50.0	50.0	56.3
lose.v	36.1	50.0	33.3	41.7	44.4	41.7	**58.3**	50.0	36.1	55.6	44.4	44.4	50.0
mean.v	60.0	65.0	65.0	60.0	57.5	65.0	70.0	**75.0**	**75.0**	67.5	**75.0**	**75.0**	**75.0**
miss.v	36.7	43.3	46.7	36.7	**50.0**	**50.0**	46.7	46.7	40.0	43.3	43.3	43.3	46.7
note.v	67.2	61.2	70.1	65.7	64.2	62.7	73.1	73.1	73.1	68.7	73.1	**76.1**	74.6
operate.v	44.4	61.1	44.4	**66.7**	44.4	44.4	**66.7**	61.1	**66.7**	55.6	61.1	61.1	**66.7**
organization.n	78.6	76.8	69.6	73.2	66.1	69.6	80.4	75.0	73.2	80.4	80.4	76.8	**82.1**
paper.n	42.7	45.3	43.6	52.1	51.3	48.7	**65.0**	51.3	53.8	62.4	60.7	59.8	**65.0**
party.n	61.2	59.5	51.7	72.4	68.1	68.1	73.3	66.4	66.4	**74.1**	69.0	70.7	70.7
performance.n	33.3	34.5	29.9	**59.8**	40.2	36.8	50.6	37.9	34.5	55.2	41.4	41.4	46.0
plan.n	75.0	73.8	72.6	81.0	78.6	77.4	86.9	83.3	77.4	84.5	86.9	84.5	**89.3**
play.v	44.2	38.5	42.3	63.5	57.7	**67.3**	65.4	57.7	63.5	61.5	61.5	59.6	63.5
produce.v	53.2	55.3	56.4	74.5	59.6	67.0	**83.0**	69.1	71.3	80.9	70.2	69.1	74.5
provide.v	82.6	89.9	87.0	91.3	85.5	88.4	**92.8**	89.9	87.0	91.3	91.3	91.3	91.3
receive.v	85.2	85.2	85.2	85.2	88.9	**92.6**	88.9	88.9	88.9	85.2	88.9	88.9	88.9
remain.v	**88.6**	84.3	84.3	75.7	85.7	77.1	84.3	85.7	87.1	82.9	85.7	87.1	85.7
rule.v	50.0	66.7	60.0	70.0	66.7	73.3	80.0	76.7	70.0	**83.3**	80.0	76.7	76.7
shelter.n	62.2	58.2	56.1	53.1	46.9	49.0	**72.4**	69.4	66.3	69.4	70.4	66.3	71.4
simple.a	22.2	**55.6**	38.9	22.2	27.8	27.8	33.3	22.2	22.2	16.7	38.9	33.3	27.8
smell.v	70.9	63.6	72.7	69.1	67.3	54.5	74.5	72.7	72.7	72.7	**76.4**	72.7	**76.4**
solid.a	17.2	6.9	17.2	17.2	**27.6**	24.1	20.7	20.7	17.2	17.2	24.1	24.1	24.1
sort.n	**72.9**	**72.9**	52.1	64.6	59.4	57.3	66.7	71.9	**72.9**	66.7	**72.9**	**72.9**	65.6
source.n	59.4	50.0	28.1	62.5	68.8	59.4	**78.1**	62.5	59.4	59.4	62.5	62.5	68.8
suspend.v	48.4	51.6	42.2	64.1	56.3	56.3	65.6	54.7	57.8	**67.2**	60.9	59.4	57.8
talk.v	69.9	69.9	**71.2**	64.4	68.5	68.5	69.9	69.9	69.9	69.9	**71.2**	**71.2**	**71.2**
treat.v	35.1	45.6	24.6	45.6	49.1	50.9	**54.4**	43.9	42.1	42.1	45.6	49.1	45.6
use.v	**100.0**	57.1	64.3	78.6	50.0	64.3	85.7	92.9	92.9	85.7	92.9	**100.0**	71.4
wash.v	58.8	55.9	61.8	64.7	55.9	**70.6**	**70.6**	64.7	64.7	61.8	**70.6**	64.7	**70.6**
watch.v	78.4	74.5	**84.3**	68.6	68.6	66.7	76.5	80.4	78.4	80.4	80.4	82.4	78.4
win.v	43.6	56.4	56.4	56.4	53.8	61.5	69.2	59.0	51.3	**71.8**	64.1	59.0	66.7
write.v	52.2	47.8	52.2	65.2	43.5	60.9	56.5	65.2	**69.6**	56.5	**69.6**	**69.6**	65.2
Average	61.9	62.5	58.6	64.1	62.4	62.3	**72.4**	68.9	66.3	70.0	70.7	70.1	71.5

on the performance of combination strategies, we conducted an experiment with those combination strategies on Pedersen's representations of context, of which the result is also shown in Table 2.

From the obtained results we see that the combination rule DS_1 based on Dempster-Shafer theory of evidence with our context representation gives the best result on average. Interestingly also, some combination strategies using OWA operators such as median and FM2 rules provide high accuracies as well. Note that though DS_1 also requires an assumption of conditional independence between individual classifiers, it seems to be reasonable since individual classifiers used here are built based on different feature sets. Furthermore, each context representation, which is used to build an individual classifier, does not provide fully enough information for detecting the sense of a target word, therefore taking weights which reflect relative confidences in individual classifiers into consideration is appropriate. This is shown by the fact that the results yielded by DS_1 rule are better than the ones obtained by DS_2 rule. It is also shown that our representation of context is much more effective than Pedersen's ones. Note that while majority voting has been widely used in many studies of combining classifiers in pattern recognition, it may not be a good choice for classifier combination in the context of WSD.

The best accuracies obtained by the DS_1 rule, 64.7% for Senseval-2 and 72.4% for Senseval-3, are comparable with the best systems in the contests for the English lexical sample tasks of Senseval-2 [13] and Senseval-3 [23], respectively. The best system of Senseval-2 contest also used a combination technique: the output of subsystems (classifiers) which were built based on different machine learning algorithms were merged by using weighted and threshold-based voting and score combination (see [35] for the detail). The best system of Senseval-3 contest used the Regularized Least Square Classification (RLSC) algorithm with a correction of the a priori frequencies (refer to [8] for more details). Note that the methods using in these systems are also corpus-based methods. The detail of this comparison is shown in Table 3.

Table 2
A comparison with Pedersen's representation of context

	Best Individual Classifier	Majority Voting	Weighted Voting	NB	DS_1	DS_2	Max	Min	Med	FM1	FM2
Ours											
Senseval-2	56.8	62.6	63.6	63.8	**64.7**	62.7	60.0	61.8	63.9	63.5	63.5
Senseval-3	64.1	69.0	70.0	71.7	**72.4**	68.9	65.8	70.2	70.6	70.0	71.5
Pedersen											
Senseval-2	57.2	59.5	60.2	55.0	59.1	**60.3**	59.6	57.2	**60.3**	60.2	60.0
Senseval-3	63.8	66.5	67.3	65.2	**68.1**	67.8	65.9	65.9	**68.1**	67.7	68.1

Table 3
A comparison with the best systems in the contests of Senseval-2 and Senseval-3

	The best system	New method – DS_1
Senseval-2	64.2%	64.7%
Senseval-3	72.9%	72.4%

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 the theories of evidence and OWA operators. By viewing distinct representations of a polysemous word as different information sources serving for identifying the right sense of the target word, we have applied Dempster rule of evidence combination to derive decision rules, denoted by DS_1 and DS_2, for making the final decision of identification. On the other hand, considering each representation of a polysemous word as information inspired by a semantics or syntactical criterion for the aim of word sense identification, we have also applied the notion of OWA operators for aggregating multi-criteria to define an overall decision function, which leads to numerous combination rules such as Max, Min, Median, FM1 and FM2, with the help of linguistic fuzzy quantifiers.

We have experimentally explored all developed combination strategies on the datasets of English lexical samples of Senseval-2 and Senseval-3. It has been shown that individual classifiers corresponding to different types of representation suitably offer complementary information about the target to be assigned a sense; it consequently makes combination strategies would help in making more correct decisions. The experimental result has shown that the discussed framework of classifier combination also yields several decision rules in WSD that perform well comparable to the best systems in the contests of Senseval-2 and Senseval-3.

For the future work, we are planning to integrate the classifier combination schemes discussed in this paper with knowledge-based WSD methods as comprehensively studied in [24] for further improving the performance of disambiguation methods.

Acknowledgements

The constructive comments and helpful suggestions from anonymous reviewers are greatly appreciated. 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] E. Agirre and P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and Applications* (Springer, Dordrecht, the Netherlands 2006).

[2] A. Al-Ani and M. Deriche, A new technique for combining multiple classifiers using the Dempster–Shafer theory of evidence, *Journal of Artificial Intelligence Research* **17** (2002) 333–361.

[3] S. Bloehdorn and H. Andreas, Text classification by boosting weak learners based on terms and concepts, *Proceedings of the fourth IEEE International Conference on Data Mining*, 2004, pp. 331–334.

[4] P. Clough and M. Stevenson, Cross-language information retrieval using Euro WordNet and word sense disambiguation, *Proceedings of Advances in Information Retrieval, 26th European Conference on IR Research (ECIR)*, 2004, Sunderland, UK, pp. 327–337.

[5] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, John A Tomlin, and Jason Y. Zien., Semtag and seeker: bootstrapping the semantic web via automated semantic annotation. *In Proceedings of the Twelfth International Conference on World Wide Web*, 2003, pp. 178–186.

[6] G. Escudero, L. Màrquez, G. Rigau, Boosting applied to word sense disambiguation, *Proceedings of the 11th European Conference on Machine Learning*, 2000, pp. 129–141.

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

[8] C. Grozea. Finding optimal parameter settings for high performance word sense disambiguation, *Proceedings of ACL/SIGLEX Senseval-3*, Barcelona, Spain, July 2004, pp. 125–128.

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

[10] V. Hoste, I. Hendrickx, W. Daelemans, A. van den Bosch, Parameter optimization for machine-learning of word sense disambiguation, *Natural Language Engineering* **8** (3) (2002) 311–325.

[11] N. Ide, J. Véronis, Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art, *Computational Linguistics* **24** (1998) 1–40.

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

[13] A. Kilgarriff, English lexical sample task description, *Proceedings of senseval-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, 2001, Toulouse, France, pp. 17–20.

[14] J. Kittler, M. Hatef, R. P. W. Duin, J. Matas, On combining classifiers, *IEEE Transactions on Pattern Anal. and Machine Intell.* **20** (3) (1998) 226–239.

[15] D. Klein, K. Toutanova, H. Tolga Ilhan, S. D. Kamvar, C. D. Manning, Combining heterogeneous classifiers for word-sense disambiguation, *ACL WSD Workshop*, 2002, pp. 74–80.

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

[17] C. A. Le, 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.

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

[19] C. A. Le, 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.

[20] C. A. Le, A. Shimazu, V.-N. Huynh, Word sense disambiguation by combining classifiers with an adaptive selection of context representation, *Journal of Natural Languague Processing* **13** (1) (2006) 75–95.

[21] C. Leacock, M. Chodorow, G. Miller, Using corpus statistics and WordNet relations for sense identification, *Computational Linguistics* **24** (1) (1998) 147–165.

[22] I.D. Melamed and P. Resnik, Tagger Evaluation Given Hierarchical Tag Sets, *Computers and the Humanities* **34** (1-2) (2000) 79–84.

[23] R. Mihalcea , T. Chklovski, A. Killgariff , The Senseval-3 English lexical sample task, *Proceedings of ACL/SIGLEX Senseval-3*, Barcelona, Spain, July 2004, pp. 25–28.

[24] A. Montoyo, A. Suarez, G. Rigau and M. Palomar, Combining knowledge- and corpus-based Word-Sense-Disambiguation methods, *Journal of Artificial Intelligence Research* **23** (2005) 299–330.

[25] R. J. Mooney, Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning, *Proceedings of the EMNLP 1996*, pp. 82–91.

[26] H. T. Ng, H. B. Lee, Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach, *Proceedings of the 34th Annual Meeting of the ACL*, 1996, pp. 40–47.

[27] T. Pedersen, A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation, *Proceedings of the North American Chapter of the ACL*, 2000, pp. 63–69.

[28] M. Sanderson, Word sense disambiguation and information retrieval, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, Dublin, Ireland, pp. 142–151.

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

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

[31] P. Vossen, G. Rigau, I. Alegria, E. Agirre, D. Farwell, M. Fuentes, Meaningful results for Information Retrieval in the MEANING project, *Proceedings of Third International WordNet Conference*, Jeju Island, Korea, 2006.

[32] X. J. Wang, Y. Matsumoto, Trajectory based word sense disambiguation, *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, August 2004, pp. 903–909.

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

[34] J. B. Yang, D. L. Xu, On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty, *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans* **32** (3) (2002) 289–304.

[35] D. Yarowsky, S. Cucerzan, R. Florian, C. Schafer, and R. Wicentowski, The Johns Hopkins SENSEVAL2 System Descriptions, *Proceedings of SENSEVAL2*, 2001, pp. 163–166.

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

[37] L. A. Zadeh, Reviews of Books: A Methematical Theory of Evidence, *The AI Magazine* **5** (1984) 81–83.