Constructing Fuzzy Relations from WordNet for Word Sense Disambiguation

Christos Diou*, Giorgos Katsikatsos and Anastasios Delopoulos
Multimedia Understanding Group,
Department of Electrical and Computer Engineering,
Aristotle University of Thessaloniki - Greece
Email: diou@mug.ee.auth.gr, gekat@ee.auth.gr, adelo@eng.auth.gr

Abstract

In this work, the problem of word sense disambiguation is formulated as a problem of imprecise associations between words and word senses in a textual context. The approach has two main parts. Initially, we consider that for each sense, a fuzzy set is given that provides the degrees of association between a number of words and the sense. An algorithm is provided that ranks the senses of a word in a text based on this information, effectively leading to word sense disambiguation. In the second part, a method based on WordNet is developed that constructs the fuzzy sets for the senses (independent of any text). Algorithms are provided that can help in both understanding and implementation of the proposed approach. Experimental results are satisfactory and show that modeling word sense disambiguation as a problem of imprecise associations is promising.

1. Introduction

Consider two sentences: "In this paper, we talk about word sense disambiguation" and "She only uses recycled paper for her notes". The word "paper" is used in both sentences but with different meanings. Word sense disambiguation is the task of mapping each word in a text to a sense (i.e., a semantic entity with specified meaning).

Word sense disambiguation is a very difficult task. In fact it is even difficult to define the senses themselves, leading several scholars to claim that there are no specified objects or concepts, that meaning depends on the linguistic use of words. At the same time, however, word sense disambiguation has attracted several scientists in the field of computational linguistics. Initially it emerged as a problem in machine translation and "traditional" AI. The difficulties of the problem, however, led researchers to abandon symbolic and deep knowledge-based approaches. Most recent works treat WSD as a problem of statistical nature, or attempt to tackle it using "shallow" knowledge-based approaches. For an excellent review of the problem of word sense disambiguation and the various approaches proposed in the bibliography, the reader is referred to [4].

The technological developments of the past decades led to the emergence of new applications where word sense disambiguation can play a critical role. Examples include

- *Information Retrieval*. Search engines can be more accurate if the search is performed based on senses and not keywords alone. This idea received criticism in [6] where the author stated that in order for WSD to be useful in information retrieval, very high precision (90% and higher) WSD algorithms must be used. In [9], however, a system was presented that exhibited increased retrieval performance with only 62.1% disambiguation accuracy.

- *Personalization*. Word sense disambiguation can be of great use to personalized text based services. More specifically, WSD can lead to efficient automatic text categorization that in turn allows the delivery of texts to users based on (semantic) profile information. Furthermore, it is of use in the task of automatically constructing profiles based on user actions since actions (e.g., the links followed in a web page) can be assigned semantic entities and not keywords.

- *Content management and annotation*. Discovering and annotating semantic interrelations between documents in large collections can be a daunting task of great cost if done manually. Word sense disambiguation can be used in exploring semantic relations to support automatic or semiautomatic content annotation and adaptation.

In the following section of this paper a method is introduced for word sense disambiguation given a fuzzy set W_{S_i} for each sense S_i. In Section 3 an approach for constructing

*Christos Diou is supported by the Greek State Scholarships Foundation

the sets W_{S_i} based on WordNet is presented. Experimental evaluation (Section 4) summarizes the results of the proposed algorithms. The paper concludes in Section 5 with some remarks and future perspectives of this work.

2. From words to word senses

2.1. Basic assumptions

Following the so-called "shallow" word sense disambiguation approach, this work does not aim at a full understanding of the text surrounding a word in order to disambiguate its sense. Instead, the goal is to model the association between the text and the word under consideration to rank its possible senses. More specifically, let a text T be a sequence of words (punctuation and syntactic structures such as sentences or paragraphs are ignored). We define the *context* C of a word w_0 at a specific position in T as a $2N+1$ word window surrounding w_0 i.e.,

$$C = [w_{-N} \dots w_0 \dots w_N]\,. \tag{1}$$

The following assumptions are made:

1. The sense of a word depends on the words appearing in its context. The part of speech, syntax and senses of these words can be ignored without great loss of information.

2. Although the exact nature of the association between a word w in C and a sense S of w_0 is unknown, it can be quantified as a degree of membership of w in the fuzzy set of words that are associated with S in a context.

The ultimate goal is to construct a fuzzy set $S_{w_0C} = d_1/S_1 + \ldots + d_n/S_n$ interpreted as the set of senses the word w_0 assumes in the context C [1]. Then, for most practical purposes, the sense of w_0 is chosen to be the one corresponding to $\max\limits_{i=1,\dots,n}(d_i)$.

2.2. Representation of senses as fuzzy sets of words

In accordance to the assumptions made in the previous section, we characterize each sense S_i by a fuzzy set

$$W_{S_i} = d_{i1}/w_{i1} + \ldots + d_{ik}/w_{ik} \tag{2}$$

of words that are associated with S_i in a textual context. Degrees of membership d_{ij} of W_{S_i} correspond to degrees of association between the words w_{ij} and S_i. A method for constructing W_{S_i} using WordNet is given in Section 3.

The weights d_{ij} can also be interpreted as the degrees of truth of the propositions $w_{ij} \Rightarrow S_i$. If we define the (crisp) variables w taking values from the set $\mathbf{W}_C$ of all words in C and S taking values from the set $\mathbf{S}_{w_0}$ of all possible senses of the word w_0 to be disambiguated, the same proposition becomes "if $w = w_{ij}$ then $S = S_i$ is true to the degree d_{ij}".

In order to evaluate the degree of truth of the proposition $S = S_i$ conjunctive inference is used on these conditionals, leading to

$$d_i = \mathop{\mathcal{U}}_{w_{ij}=w_l}(d_{ij}) \tag{3}$$

for $i = 1, \dots, n$ (all senses of w_0), $j = 1, \dots, k$ (all words used to define W_{S_i}), $l = -N, \dots, N$ and $w_l \in C$ (all words in the context) where $\mathcal{U}$ is a fuzzy union operator. Note $\mathcal{U}$ can be any fuzzy t-conorm (e.g., the algebraic sum $\mathcal{U}(a,b) = a + b - ab$ and not just the standard fuzzy union i.e., $\max$). The choice of $\mathcal{U}$ depends on whether a super-idempotent t-conorm ($\mathcal{U}(a,a) > a$) is desirable. Super-idempotency enables the increase of d_i if a word $w_{ij} \in W_{S_i}$ appears more than once in C. The experimental results presented in Section 4 were obtained using the algebraic sum operator, although the method proved to be rather insensitive to the choice of the union type.

The method presented above leads to the construction of the set S_{w_0C} and is summarized in Algorithm 1.

Algorithm 1 Sense ranking.

Input: n possible senses S_i associated with word w_0 to be disambiguated, a fuzzy set W_{S_i} for each sense and the context C.
Output: The membership degrees d_i of S_i in the fuzzy set S_{w_0C}.

for $i := 1$ to n **do**
    $d_i := 0$
    for $l := -N$ to N **do**
        $j :=$ `search_word`(W_{S_i}, w_l)
        if $j > 0$ **then**
            $d_i = \mathcal{U}(d_i, d_{ij})$
        end if
    end for
end for

Notes: `search_word` is an arbitrary search algorithm that returns the positive index j of the word w_{ij} if $w_l = w_{ij}$ (i.e., if w_l is found in the set W_{S_i}) or negative otherwise.

Since C is readily provided from the given text document, the inputs to Algorithm 1 that remain to be defined are the senses S_i and the fuzzy sets W_{S_i}, $i = 1, \dots, n$ characterizing each sense. The method proposed in this work relies on WordNet to obtain this information.

[1] The fuzzy set S_{w_0C} can also be seen as a fuzzy restriction and thus as a possibility distribution where $\pi(S_i) = d_i,\ i = 1, \dots, n$ is the possibility of w_0 in C to assume sense S_i. Note, however, that the approach used to derive the weights d_i does not utilize possibility theory.

3. WordNet as a fuzzy knowledge base

WordNet is a lexical database that generally provides the following types of information:

1. Senses, corresponding to conceptual entities. A sense is defined by a set of synonymous words (*synset*).
2. Relations between senses e.g., the hyponymy ("is a kind of") and the meronymy ("has part") relations.
3. Additional information about senses, such as a small description of each sense (*gloss*), the estimated frequency of appearance of a sense with respect to a word etc.

The reader is directed to [3] and [10] for more information.

The senses S_i, $i = 1, \dots, n$ of a word are obtained through WordNet. In order to define W_{S_i} for each sense we utilize the relations between senses, the synset of each sense and optionally frequency information.

3.1. Extracting fuzzy relations

Relations in WordNet are crisp. There is no imprecision whatsoever associated with the statement "*allergology#1 is a kind of medicine #1*"[2]. However, the existence of a relation $R_t(S_j, S_i)$ of type t (e.g., hyponymy) between senses S_j and S_i does not necessarily imply that words in the synset of S_j will appear in the context of a word from the synset of S_i. We are compelled to make an additional assumption:

3. A sense S_j has a degree of membership to a fuzzy set of senses that are associated with S_i in a context if the two senses are related in WordNet with $R_t(S_j, S_i)$.

For instance, consider sense 1 of the word "bank" and one of its hyponyms, "credit union" (which is monosemous). We assume that "credit union#1" belongs to the set of senses associated with "bank#1" in a context up to a degree, while the sense "riverside#1", a hyponym of "bank#2", is not.

Observe that in the absence of additional information, the aforementioned degree of membership d_t depends on the type t of relation R_t. In our experiments we utilized hyponymy, hyperonymy, holonymy, meronymy and domain relations with weights 0.9, 0.4, 0.4, 0.9 and 0.9 respectively. These weights are subjective evaluations signifying the importance of each relation and were not derived experimentally. In summary, for each relation $R_t(S_j, S_i)$ in WordNet a fuzzy relation $\mathbf{R}_t(S_j, S_i) = d_t$ is constructed where d_t is the weight assigned to R_t.

[2] $w\#n$ reads: Word w used with its n'th sense

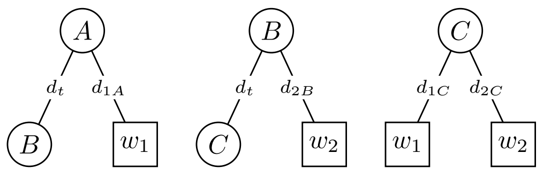

Figure 1. Graphical representation of an example R_t. Senses are noted in circles while square boxes indicate words from the synsets.

3.2. Quantifying word - sense relations

Relations R_t and consequently $\mathbf{R}_t$ are not transitive. If $(A, B) \in R_t$ and $(B, C) \in R_t$ then $(A, C) \notin R_t$ in general. Of course, for the transitive closure R_t^+ of R_t, $(A, C) \in R_t^+$. The same is true for the fuzzy relations $\mathbf{R}_t$, where the transitive closure is obtained via the $\sup$-t composition ($\max$-t in our case, since relations are discrete).

Consider Figure 1 and assume that the degrees d_{iS} between a word and a sense are known. We extend $\mathbf{R}_t$ so as to include this word-sense information as well and calculate its transitive closure which gives $\mathbf{R}_t^+(w_1, A) = \max(d_{1A}, \mathcal{I}(d_{1C}, d_t, d_t))$ and $\mathbf{R}_t^+(w_2, A) = \max(\mathcal{I}(d_{2B}, d_t), \mathcal{I}(d_{2C}, d_t, d_t))$ where $\mathcal{I}$ is a fuzzy t-norm (e.g., algebraic product $\mathcal{I}(a, b) = ab$). Once again the properties of the standard union may not be desirable, since one may want $\mathbf{R}_t^+(w_1, A) = \mathcal{U}(d_{1A}, \mathcal{I}(d_{1C}, d_t, d_t)) \geq \max(d_{1A}, \mathcal{I}(d_{1C}, d_t, d_t))$ (see [5], p. 78 for a proof). It is possible to use any union operator $\mathcal{U}$ instead of $\max$ in the above equations, but certain issues arise: (i) There is no transitive closure and from an algorithmic point of view one must define the number of compositions applied for the calculation. (ii) Generally, for $\mathbf{R}_t^+(w_2, A)$ in Figure 1 $\mathcal{U}(\mathcal{I}(d_{2B}, d_t), \mathcal{I}(d_{2C}, d_t, d_t)) \neq \mathcal{I}(\mathcal{U}(\mathcal{I}(d_{1C}, d_t), d_{2B}), d_t)$ (for $\max$ the equality holds) and one of the two ways of applying the composition must be selected (since the order of operations is now important, distributivity does not hold).

The last piece of information that is required is the degree up to which a word w in the synset of S_i is associated with S_i (e.g., d_{1A} in Figure 1). One way of automatically computing this number is by exploiting the frequency score of S_i. If a word has n possible senses we assign a higher membership degree to the more frequent senses. In our experiments, a sigmoid-like function was used

$$d_{wS_i} = 0.3 + 0.7 \frac{1}{1 + e^{-2(x-c)}} \tag{4}$$

where x is the number of occurrences of S_i (frequency) given by WordNet and $c = f_{\max} - f_{\min}$ is the difference of occurrence between the most and least frequent senses of w.

The elements needed to construct W_{S_i} are obtained by $\mathbf{R}_t^+(w_{ij}, S_i) = d_{ij}$ for all relations t and words w that have a non-zero membership degree. The process is presented in a form suitable for implementations in Algorithms 2 and 3 where for the benefit of computational complexity a limited depth search is applied for all relations. This is due to the observation that relations that have a depth greater than 3 seem to have little effect on the disambiguation results for the experiments that were performed. Note that these algorithms do not depend on the context C or the word to be disambiguated and can be applied a priori to construct a fuzzy knowledge base for word sense disambiguation based on WordNet.

As a final post processing step, all degrees in d_i of S_{w_0C} were reduced by a penalty $d'_i = d_i - \text{length}(W_{S_i})/1000$. This is for normalization purposes, since for a W_{S_i} with a large number of elements, the probability of matching a word in C (and thus increasing the degree d_i is higher).

Algorithm 2 Depth first tree traversal of relation $\mathbf{R}_t$ and retrieval of the elements of W_{S_i} (or equivalently calculate $\mathbf{R}_t^+(w, S_i)$ for all words with significant membership degree).

Input: WordNet information, W_{S_i}, t, the current sense S the current depth of the traversal and a depth threshold T.
Output: Updated W_{S_i}.

if depth $:= 0$ **then**
  $S :=$ next sense in depth first traversal
  depth $:= 1$
  W_{S_i} :=`traverse_tree`(W_{S_i}, t, S, depth, T)
  return
end if
if depth $> T$ **then**
  return
end if
for all elements w in the synset of S **do**
  d_f := (given by Equation (4) if frequency is taken into account or 1 otherwise)
  $d_w := \mathcal{I}(\underbrace{d_t, \ldots, d_t}_{\text{depth times}}, d_f)$
  if w already in W_{S_i} with d'_w **then**
    $d_w := \mathcal{U}(d'_w, d_w)$
    Change membership degree to d_w
  else
    Add w to W_{S_i} with degree d_w
  end if
end for
$S :=$ next sense in depth first traversal
depth $:=$ depth $+ 1$
W_{S_i} :=`traverse_tree`(W_{S_i}, t, S, depth, T)

Algorithm 3 Calculate all W_{S_i}

Input: WordNet information, the n senses of the word w_0 to be disambiguated and the traversal depth threshold T.
Output: W_{S_i}, $i = 1, \ldots, n$.

for $i := 1$ to n **do**
  /* Initialize and get synset of S_i. */
  $W_{S_i} = \emptyset$
  for all elements w in the synset of S_i **do**
    Add w to W_{S_i} with degree given by (4) or 1 if frequency is not taken into account
  end for
  for all types of relations t **do**
    /* `traverse_tree` is given in Algorithm 2 */
    depth $:= 0$
    W_{S_i} :=`traverse_tree`(W_{S_i}, t, S_i, depth, T)
  end for
end for

4. Experiments

4.1. Experimental setup

Evaluation was performed using SemCor 2.0, a collection of documents from the Brown Corpus. More specifically, the method was tested on 103 sense tagged texts (brown1 collection of SemCor). The results below will be in aggregated form for all evaluation data. Automatic sense tagging was performed on *all* tagged nouns in SemCor 2.0 and only nouns were taken into account in the context C as well.

A note must be made about fine versus coarse grained disambiguation. In WordNet, it is common for senses of the same word to overlap i.e., to have similar but not exactly the same meaning. Word sense disambiguation of fine-grained senses is a very difficult task and this is depicted by the fact that usually the human interannotator agreement is low (72.5% for 2,212 words in Senseval-3 [7]) limiting the upper bounds of the expected machine performance.

The adjustable parameters of the proposed method are (i) the choice of fuzzy operators ($\mathcal{U}$ and $\mathcal{I}$), (ii) the length of the context window (N) (iii) the degrees d_t assigned to the various types of WordNet relations and (iv) the depth T of the search. Apart from extreme cases, the results were generally not significantly affected by modifications of these values. The values used in the following experiments are shown in Table 4.1.

4.2. Results

The tables below indicate the mean, standard deviation, best and worse precision results for all the test data (for polysemous words only and for all words). These results were

Parameter	Value
$\mathcal{U}(a,b)$	$a+b-ab$
$\mathcal{I}(a,b)$	ab
N	36
d_t	(see Section 3.1)
T	4

Table 1. Parameter values used in the experiments.

Measure	*Precision* (%) (polysemous)	*Precision* (%) (all words)
Mean	58.5	67.2
σ	6.6	6.4
max	80	83.6
min	43.9	49.7

Table 2. Results without frequency information for 103 texts.

obtained from two variants of the method, one that utilizes frequency information and one that does not.

4.3. Comparison with related work

The bibliography in the wider area of word sense disambiguation is vast and the various approaches differ in their information sources. Systems that employ supervised and unsupervised training in conjunction with probability theory have received much attention, while others propose the combination of information sources towards WSD [8]. The method proposed in this paper uses a single source (WordNet) and no training is performed, an advantage in real world applications.

The best system in the all-words (not only nouns) task of senseval-3 achieved 65.2% precision. Although the results presented in this paper correspond to noun disambiguation only, the results show that proposed method seems promising in terms of performance.

For a more in-depth comparison of our algorithm with related works we implemented a method proposed by Agirre and Rigau [1] that seems to fall in the same category (since it is solely based on WordNet information). In this work, the authors select four texts from the brown corpus collection and evaluate word sense disambiguation with a metric called "conceptual density" that is calculated based on WordNet hyponymy. No frequency information is utilized. They report the system's peak precision to be at 64.5 for both monosemous and polysemous nouns.

Measure	*Precision* (%) (polysemous)	*Precision* (%) (all words)
Mean	62.7	70.5
σ	6.6	6.0
max	81.4	84.8
min	44.3	56.6

Table 3. Results with frequency information for 103 texts.

We implemented the aforementioned method in order to test its performance on the entire brown1 collection. The mean precision derived in our experiments was 33.7%, the standard deviation 8.1%, the minimum precision achieved was 16.7% and the maximum 62.9%. These lower numbers can be attributed to two main reasons: (i) We did not experiment enough with conceptual density in order to derive the optimal set of parameters hence the method surely can provide better results and (ii) in [1] the results are reported as pairs of precision and recall, meaning that they are evaluated on a subset of nouns on the text while we provide results based on all tagged nouns of the test collection without discriminating between successful and unsuccessful disambiguation. Ultimately this can lead to reduction of the reported precision levels, but can make the two methods directly comparable.

5. Discussion

In this work the problem of word sense disambiguation was formulated as a problem of imprecision. Several assumptions were made that allowed the use of WordNet to construct a fuzzy knowledge base for WSD. The results were satisfactory, but the authors believe that the only indication that this provides is that the presented approach is promising. The problem is far from being solved and several open issues can be identified:

1. Evaluate the use of other information sources apart from WordNet (e.g., collocations) with the inference method presented in Section 2.

2. Measure the role of fine-grained senses in the disambiguation process. What would be the improvement of results if a coarse grained lexical database was used?

3. Derive a similar formulation that would be based on uncertainty and belief measures [2] (e.g., using possibility theory [11]) and compare the results.

4. Apply the method in practical scenarios and evaluate its effectiveness in real world problems.

The last point is especially important since specific application settings or knowledge domains may allow for improvements in WSD algorithms in terms of performance and above all, in terms of practical use.

References

[1] E. Agirre and G. Rigau. Word sense disambiguation using conceptual density. In *COLING '96*, pages 16–22, Copenhagen, Danmark, 1996.

[2] D. Dubois and H. Prade. Possibility theory, probability theory and multiple-valued logics: A clarification. *Annals of Mathematics and Artificial Intelligence*, 32:35–66, 2001.

[3] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, May 1998.

[4] N. Ide and J. Veronis. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40, 1998.

[5] G. J. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic; Theory and Applications*. Prentice Hall, 1995.

[6] M. Sanderson. Word sense disambiguation and information retrieval. In *SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, pages 49–57, Dublin, IE, 1994.

[7] SENSEVAL-3, 3rd ACL-SIGLEX organized Evaluation Excercises for the Semantic Analysis of Text, homepage `http://www.senseval.org/senseval3`, 2004.

[8] M. Stevenson and Y. Wilks. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3):321–349, 2001.

[9] C. Stokoe, M. P. Oakes, and J. Tait. Word sense disambiguation and information retrieval revisited. In *SIGIR-03, 26th ACM International Conference on Research and Development in Information Retrieval*, pages 159–166, Toronto, Canada, 2003.

[10] WordNet, a lexical database for the English language, homepage `http://wordnet.princeton.edu`.

[11] L. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.