A Proposal for WSD using Semantic Similarity

Susana Soler and Andrés Montoyo

Research Group of Language Processing and Information Systems
Department of Software and Computing Systems
University of Alicante, Alicante, Spain
{susana, montoyo}@dlsi.ua.es

Abstract. The aim of this paper is to describe a new method for the automatic resolution of lexical ambiguity of verbs in English texts, based on the idea of semantic similarity between nouns using WordNet.

1 An outline of our approach.

The method of WSD proposed in this paper is based on knowledge and consists basically of sense-disambiguating of the verb that appear in an English sentence.
A simple sentence or question can usually be briefly described by an action and an object [1]. For example the main idea from the sentence "*He eats bananas"* can be described by the action-object pair *"eat-banana"*. Our method determine which senses of these two words are more similar between themselves.
For this task we use the concept of semantic similarity [2] between nouns based on WordNet [3] hierarchy. In WordNet, the gloss of a verb synset provides a noun-context for that verb, i.e. the possible nouns occurring in the context of that particular verb [1]. The glosses are used here in the same way a corpus is used.
Our method takes into consideration the verb-noun pair extracted from the sentence. This verb-noun pair is the input for the algorithm. The output will be the sense tagged verb-noun pair, so we assign the sense of the verb. The algorithm is described as follows:
Step 1. Determine all the possible senses for the verb and the noun by using WordNet. Let us denote them by $<v_1, v_2, \ldots, v_k>$ and $<n_1, n_2, \ldots, n_m>$
Step 2. For each sense of verb v_h and all senses of noun $<n_1, n_2, \ldots, n_m>$:
2.1. Extract all the glosses from the sub-hierarchy including v_h. The sub-hierarchy including a verb v_h is determined as follows: consider the hypernym h_h of the verb v_h and consider the hierarchy having h_h as top [1].
2.2. Determine the nouns from these glosses. These constitute the noun-context of the verb. Determine all the possible senses for all these nouns. Let us denote them by $<x_1, x_2, \ldots, x_n>$.
2.3. Then we obtain the similarity matrix (Sm) using the semantic similarity, where each element is defined as follows:

$$Sm(i, j) = sim (x_i, n_j)$$

For determining the semantic similarity (sim(x_i, n_j)) between each sense of the nouns extracted from the gloss of verb and each sense of the input noun, we use the formula followed:

$$\text{sim}\ (x_i, n_j) = 1 - \text{sd}\ (x_i, n_j)^2$$

$$\underline{\text{sd}}\ (x_i, n_j) = \frac{1}{2} . \left(\frac{D1 ? D}{D1} + \frac{D2 ? D}{D2} \right)$$

where sim (x_i, n_j) is the semantic similarity between two concepts defined by their WordNet synsets x_i and n_j; sd (x_i, n_j) is the semantic distance for nouns. D1 is the depth of synset x_i, D2 is the depth of synset n_j, and D is the depth of their nearest common ancestor in the WordNet hierarchy.

2.4. Determine the total similarity between the sense *h* of verb (v_h) and all the senses of input noun <n_1, n_2, …, n_m>. For each n_j:

$$\text{Ts}(h, j) = \sum_{i?1}^{n} \text{sim}\ (x_i, n_j)$$

where n is the number of nouns extracted from the gloss of the sense *h* of the verb.

Step 3

To resume all similarity matrixes (Sm) obtained in step2 for each sense of verb, we make now the total similarity matrix (Tsm) composed by total similarity (Ts) for each sense of verb and each sense of noun. Each element of this matrix is defined as follows:

$$\text{Tsm}\ (i, j) = \text{Ts}\ (i, j)$$

Step 4

The most similar sense combination scores the highest value in the total similarity matrix (Tsm). So the output of the algorithm is the pair verb-noun (v_i-n_j) that contains this value in the matrix. Therefore the sense of the verb is chosen and given as the solution.

Consider as an example of a verb-noun pair the phrase *rewrite-article* extracted from the sentence "*She rewrites the article once again*". The verb *rewrite* has two senses and the noun *article* has four senses in WordNet version 1.5.

From the sense1 of verb *rewrite* we extract the nouns from its gloss. Then we have <student, thesis, week>. We obtain the semantic similarity matrix (Sm1).

rewrite1	article1	article2	article3	article4
student1	0.31	0.37	0	0
student2	0.45	0.40	0	0
thesis1	0.67	0	0.70	0.40
thesis2	0.72	0	0.94	0.44
week1	0.29	0	0.30	0.30
week2	0.29	0	0.30	0.30
week2	0.26	0	0.27	0.27
Ts1	**2.99**	**0.77**	**2.51**	**1.71**

From the sense2 of verb *rewrite* we extract the nouns from its gloss: <purpose, play, schools, work, poem, novels>. We would obtain the following total similarity (Ts).

rewrite2	article1	article2	article3	article4
Ts2	**2.84**	**0.83**	**2.45**	**1.46**

We obtain the total similarity matrix (Tsm):

Tsm	article1	article2	article3	article4
Rewrite1	**2.99**	0.77	2.51	1.71
Rewrite2	2.84	0.83	2.45	1.46

The most similar sense combination is the sense one of the noun *article* and the sense one of the verb *rewrite*. So the output of the algorithm is the pair verb-noun: *rewrite1-article1* that contains the highest value in the matrix. The sense one of the verb *rewrite* is chosen as the solution.

3 Conclusion and Further Work

In this paper, we have presented a method for WSD that is based on semantic similarity between nouns using WordNet. Although this method has been presented as stand-alone, it is our belief that our method could be combined with other methods or could be a new heuristic of another method. In further work we intend to modify the method by adding more lexical categories for disambiguating adjectives and adverbs using the gloss of a noun synset. Finally, we pretend to test this method on sentences taken from Semcor.

References

1. Mihalcea R. and Moldovan D. (1999) *A Method for word sense disambiguation of unrestricted text*. Proc. 37th Annual Meeting of the ACL 152-158, Maryland, Usa.
2. Stetina J., Kurohashi S. and Nagao M. (1998) *General word sense disambiguation method based on full sentencial context*. In Usage of WordNet in Natural Language Processing. COLING-ACL Workshop, Montreal, Canada.
3. Miller G.A. (1990) *WordNet: An on-line lexical database.* International Journal of Lexicography, 3(4): 235-312.