JIGSAW: UN ALGORITMO PER IL WORD SENSE DISAMBIGUATION

JIGSAW: AN ALGORITHM FOR WORD SENSE DISAMBIGUATION

PIERPAOLO BASILE · GIOVANNI SEMERARO

SOMMARIO/*ABSTRACT*

In questo articolo descriveremo JIGSAW, un algoritmo per WSD di tipo *knoweldge-based* (basato su conoscenza) che cerca di disambiguare tutte le parole presenti in un testo utilizzando una risorsa lessicale. L'assunzione che sta alla base dell'algoritmo è che una strategia specifica per ogni parte del discorso (Part-Of-Speech) è più efficiente di una singola strategia.

Word Sense Disambiguation (WSD) is traditionally considered a complex task. Advances in this field would have a significant impact on many relevant web-based applications, such as information retrieval and information extraction. This paper describes JIGSAW, a knowledge-based WSD system that attempts to disambiguate all words in a text by exploiting external lessical knowledge-base. The main assumption is that a specific strategy for each Part-Of-Speech (POS) is better than a single strategy. We evaluated the accuracy of JIGSAW on EVALITA Word Sense Disambiguation All-Word-Task, which consists in tagging almost all of the content words within a corpus containing about 5000 words extracted from the Italian Syntactic Semantic Treebank. Nouns, verbs, adjectives and a small set of proper nouns are semantically tagged with senses obtained by ItalWordNet.

Keywords: Natural Language Processing, Word Sense Disambiguation

1 JIGSAW

The goal of a WSD algorithm consists in assigning a word w_i occurring in a document d with its appropriate meaning or sense s, by exploiting the *context* C in where w_i is found. The context C for w_i is defined as a set of words that precede and follow w_i. The sense s is selected from a predefined set of possibilities, usually known as *sense inventory*. In the proposed algorithm, the sense inventory is obtained from ItalWordNet, according to EVALITA All-word-Task instructions. JIGSAW is a WSD algorithm based on the idea of combining three different strategies to disambiguate nouns, verbs, adjectives and adverbs. The main motivation behind our approach is that the effectiveness of a WSD algorithm is strongly influenced by the POS tag of the target word. An adaptation of Lesk dictionary-based WSD algorithm has been used to disambiguate adjectives and adverbs [1], an adaptation of the Resnik algorithm has been used to disambiguate nouns [4], while the algorithm we developed for disambiguating verbs exploits the nouns in the *context* of the verb as well as the nouns both in the glosses and in the phrases that ItalWordNet utilizes to describe the usage of a verb. JIGSAW takes as input a document $d = (w_1, w_2, \ldots, w_h)$ and returns a list of ItalWordNet synsets $X = (s_1, s_2, \ldots, s_k)$ in which each element s_i is obtained by disambiguating the *target word* w_i based on the information obtained from ItalWordNet about a few immediately surrounding words. We define the *context* C of the target word to be a window of n words to the left and another n words to the right, for a total of $2n$ surrounding words. The algorithm is based on three different procedures for nouns, verbs, adverbs and adjectives, called $JIGSAW_{nouns}$, $JIGSAW_{verbs}$, $JIGSAW_{others}$, respectively. Follows a short description of procedures $JIGSAW_{nouns}$ and $JIGSAW_{verbs}$, whereas the detailed description for all procedure is in [2].

1.1 $JIGSAW_{nouns}$

The procedure is obtained by making some variations to the algorithm designed by Resnik for disambiguating noun groups. Given a set of nouns $W = \{w_1, w_2, \ldots, w_n\}$, obtained from document d, with each w_i having an associated sense inventory $S_i = \{s_{i1}, s_{i2}, \ldots, s_{ik}\}$ of possible senses, the goal is assigning each w_i with the most appropriate sense $s_{ih} \in S_i$, according to the *similarity* of w_i with the other words in W (the context for w_i). The idea is to define a function $\varphi(w_i, s_{ij})$, $w_i \in W$,

$s_{ij} \in S_i$, that computes a value in $[0, 1]$ representing the confidence with which word w_i can be assigned with sense s_{ij}. $JIGSAW_{nouns}$ differs from the original algorithm by Resnik in several ways. First, in order to measure the relatedness of two words we adopted a modified version of the Leacock-Chodorow measure [3], which computes the length of the path between two concepts in a hierarchy by passing through their *Most Specific Subsumer* (MSS). In our version, we introduced a constant factor $depth$ which limits the search for the MSS to $depth$ ancestors, in order to avoid "poorly informative MSSs". Moreover, in the similarity computation, we introduced both a Gaussian factor $G(pos(w_i), pos(w_j))$, which takes into account the distance between the position of the words in the text to be disambiguated, and a factor $R(k)$, which assigns s_{ik} with a numerical value, according to the frequency score in ItalWordNet (more importance is given to the synsets that are more common than others).

1.2 $JIGSAW_{verbs}$

We define the *description* of a synset as the string obtained by concatenating the gloss and the sentences that ItalWordNet uses to explain the usage of a synset. First, $JIGSAW_{verbs}$ includes, in the context C for the target verb w_i, all the nouns in the window of $2n$ words surrounding w_i. For each candidate synset s_{ik} of w_i, the algorithm computes $nouns(i, k)$, that is the set of nouns in the description for s_{ik}. Then, for each w_j in C and each synset s_{ik}, the following value is computed:

$$max_{jk} = max_{w_l \in nouns(i,k)} \{\mathtt{sim}(w_j, w_l, depth)\} \tag{1}$$

where $\mathtt{sim}(w_j, w_l, depth)$ is the same similarity measure adopted by $JIGSAWnouns$. In other words, max_{jk} is the highest similarity value for w_j wrt the nouns related to the k-th sense for w_i. Finally, an overall similarity score among s_{ik} and the whole context C is computed:

$$\varphi(i, k) = R(k) \cdot \frac{\sum_{w_j \in C} G(pos(w_i), pos(w_j)) \cdot max_{jk}}{\sum_h G(pos(w_i), pos(w_h))} \tag{2}$$

where both $R(k)$ and $G(pos(w_i), pos(w_j))$, that gives a higher weight to words closer to the target word, are defined as in $JIGSAWnouns$. The synset assigned to w_i is the one with the highest φ value.

2 Experiments

$JIGSAW$ is implemented in JAVA. Experiments were performed by using the instructions for EVALITA WSD All-Word-Task. The dataset consisted of about 5000 words. Precision and Recall are reported in Table 1.

The results are encouraging as regards precision, considering that our system exploits only ItalWordNet as knowledge base. JIGSAW was compared only with the *baseline* (for all words, the first sense in ItalWordNet is selected), which achieves very high results. It is important to notice that the process of WSD requires a pre-processing phase, that includes lemmatization and POS-tagging, which introduce errors influencing the recall. We have estimated the lemmatization and POS tagging precision respectively to 77,66% and 76,23%.

$SYSTEM$	P	R	*attempted*
$UniBa\_Basile$ *(JIGSAW)*	0.560	0.414	73.95%
1*st sense (baseline)*	0.669	0.669	100%

Table 1: JIGSAW results on EVALITA All-Words Task

REFERENCES

[1] S. Banerjee and T. Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *CICLing'02: Proc. 3rd Int'l Conf. on Computational Linguistics and Intelligent Text Processing*, pages 136–145, London, UK, 2002. Springer-Verlag.

[2] P. Basile, M. de Gemmis, A.L. Gentile, P. Lops, and G. Semeraro. Jigsaw algorithm for word sense disambiguation. In *SemEval-2007: 4th International Workshop on Semantic Evaluations*, pages 398–401. ACL press, 2007.

[3] C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. In *C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database*, pages 305–332. MIT Press, 1998.

[4] P. Resnik. Disambiguating noun groupings with respect to WordNet senses. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 54–68. Association for Computational Linguistics, 1995.

CONTACTS

PIERPAOLO BASILE, GIOVANNI SEMERARO
Dipartimento di Informatica, Università di Bari
Via E. Orabona, 4
70125 Bari
Email: {basilepp | semeraro}@di.uniba.it

PIERPAOLO BASILE is a PhD student in Computer Science at the University of Bari. His research is devoted to the application of machine learning techniques for NLP. The topic of his studies concerns word sense disambiguation and semantic text analysis obtained by integrating lexical resources, ontologies and statistical methods in the learning process in order to extract concepts from unstructured text.