A Word Sense Disambiguation Method for Semi-automatic Conversion of NL Text into CDL (Concept Description Language)

Francisco A Tacoa R[†] Hiroshi Uchida[‡] Mitsuru Ishizuka[†]

Abstract

CDL is a language to represent semantic meaning of contents in a simple and structured manner. In particular, it is intended to represent the meanings of Web texts so that computers can understand and manipulate them. In this paper, we present a WSD (Word Sense Disambiguation) method for selection of best candidates of word meanings. This method works inside an application that performs semi-automatic conversion from English NL text to CDL format.

1 Introduction

Word Sense Disambiguation (WSD) is the process of selecting the correct sense for a word with multiple meanings. It is a very complex and important task inside Natural Language Processing, especially when using Information Retrieval and Machine Translation systems, among others.

1.1 Definition of CDL

CDL is a computer language for describing the concept structure of contents, as explained in [3]. Some of its roles as a language are:

- to represent semantic meaning of texts,
- to overcome language barriers,
- to realize machine understandability.

CDL includes two basic elements that compose the whole conceptual structure:

- “Entity”, to indicate concepts;
- “Relation”, to indicate a link between two concepts.

This language can be represented in two formats: Text Notation and Graph Notation.

2 Proposal

The WSD method we are proposing runs inside an application called CDL Graphical Editor. This application serves for testing the method accuracy, as well as for textual and graphical representation of the CDL structure.

For the WSD process, CDL Graphical Editor uses words information from Universal Network Language Knowledge Base (UNLKB) [1] as the unique data source. UNLKB is a semantic ontology tree, where concepts exist under logical constraints. Words senses inside UNLKB are referred to as concepts.

A noun concept contains information about its parent concept, and a verb concept contains also information about word classes, which are connected to that verb concept through a semantic relation label, such as “agt” (Agent), “obj” (Object), etc.

In order to perform WSD, the application analyzes words co-occurrences through the following steps:

- Extraction of syntactic analysis from NL text. In this step, we use the RelEx Dependency Relation Extractor tool [2].
- Extraction of words candidates from UNLKB.
- Conversion of syntactic relations to semantic relations. A very basic rule-based method is used for this purpose.
  Calculation of best candidates for words, based on semantic distance. Candidates of related words from sentences. In the case of verbs candidates, best candidates will be those ones with the biggest quantity of semantic relations connected with nouns candidates; and in the case of nouns candidates, best candidates will be determined by the closest distance to a word

[†] Graduate School of Information Science and Technology, The University of Tokyo

[‡] UNDL Foundation

class connected to the respective word class.

Currently, the method is limited to the analysis of nouns and verbs.

3 Experiments and Results

In order to test our method, we selected a total of 20 sentences that contain words taken from UNLKB. The sentences contain one verb and up to four nouns. Other word types are ignored. For each sentences, the values that were calculated are: Total Ambiguous Words (TAWs); Total Correct Selections (TCSs); and Accuracy (Acc). Accuracy indicates the percentage of precision for the method:

$$\mathrm{Acc_i} = \frac{\mathrm{TCSs_i}}{\mathrm{TAWs_i}} \qquad (1)$$

where *i* is the index of the sentence.

Table 1. Experiment Results

	TAWs	TCSs	Acc
S_1	2	2	1
S_2	3	3	1
S_{14}	2	1	0.5
S_{20}	3	2	0.6667

In this table, values for sentences 1, 2, 14, and 20 are the most representative of the whole experiment.

When all 20 sentences are taken into account, the method reaches a total of 83.33%

4 Discussion

The WSD method relies deeply in the way that concepts are organized inside UNLKB. For instance, the probability of a word candidate to be selected as best candidate may vary according to its distance value respect to another word candidate.

The data available from UNLKB that was used for the experiment contains about 980 words, for which exist near 2000 concepts. However, the sentence list could not be constructed easily.

5 Conclusions and Future Work

For the method presented in this paper, we can conclude the following:

- WSD task requires a very complex analysis of sentences, in order to get correct words senses.
- The employment of UNLKB data makes possible the calculation of distance values for word senses.
- Results may improve if the best candidate selection method works along with additional data, such as context.

Our future work will be focused on the improvement of the WSD method, by including logical analysis of words co-occurrences with statistical data.

6 References

[1] UNDL Foundation. UNL (Universal Networking Language). [Online]. http://www.undl.org/

[2] OpenCog. RelEx Dependency Relationship Extractor. [Online]. http://opencog.org/wiki/RelEx

[3] T. Yokoi, H. Yasuhara, H. Uchida, Zhu Meiying, and K. Hasida, "CDL (Concept Description Language): A Common Language for Semantic Computing," in *WWW2005 Workshop on the Semantic Computing Initiative (SeC2005)*, Makuhari, Japan, 2005.