A Word Sense Disambiguation Approach for the Conversion of NL Text to Concept Description

Francisco Tacoa[*1] Hiroshi Uchida[*2] Mitsuru Ishizuka[*1]

[*1] Graduate School of Information Science and Technology, The University of Tokyo
[*2] UNDL Foundation

Concept Description Language (CDL) is a language to represent semantic meaning of web contents so that computers can understand and manipulate them. One problem is that some texts have multiple meanings. Hence, it is necessary to perform selection of word meanings prior to conversion. As a possible solution, this paper presents a method for selection of best candidates for word meanings.

1. Introduction

Word Sense Disambiguation (WSD) is the process of selecting the most appropriate meaning for a word with multiple senses. It is an intermediate task that allows the correct execution of others, such as Machine Translation (MT), Information Retrieval (IR) and Information Extraction (IE).

There are so many works related to word senses that it is practically impossible to describe them all. Most of them can be referred from [Agirre 2006].

1.1 Concept Description Language

According to [Yokoi 2005], CDL is an artificial language that describes the conceptual structure of contents. Some of its purposes are: (1) to represent semantic meaning of texts; (2) to overcome language barriers, and (3) to realize machine understandability.

CDL contains two basic elements that construct the whole conceptual structure:

- "Entities", to indicate concepts;
- "Relations", to indicate links between two concepts.

Examples of notations for entities and relations can be detailed in Figure 1 and Figure 2, respectively:

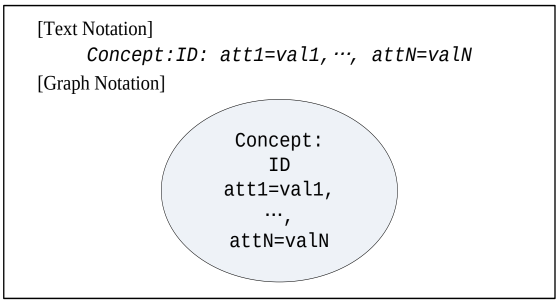


Figure 1. Entities in CDL

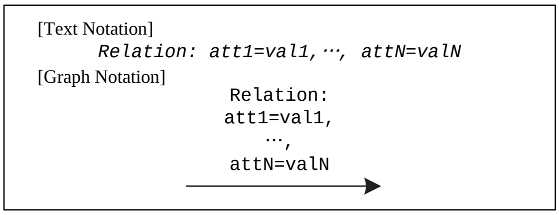


Figure 2. Relations in CDL

There are two ways to represent CDL: Text Notation and Graph Notation. Text Notation consists of a structured description where Entities and Relations are differentiated by symbolic texts.

The Graph Notation is defined as a directed graph, where Entities become nodes and Relations become arcs connecting two nodes. Entities may also contain complex structure within, that is, an inner network structure. In such cases, they are referred to as hyper-nodes.

CDL is divided into several languages:

- CDL.core: constitutes the basis for all CDL family languages,
- CDL.nl: part of CDL intended for representation of semantics in natural language texts,
- CDL.unl: part of CDL for representation of the Universal Networking Language (UNL) [Uchida 2005],
- CDL.math, CDL.prog, CDL.movie, CDL.music, etc: different CDL specifications for each corresponding language or media.

Since our work is related to natural language, we will refer only to CDL.nl hereafter.

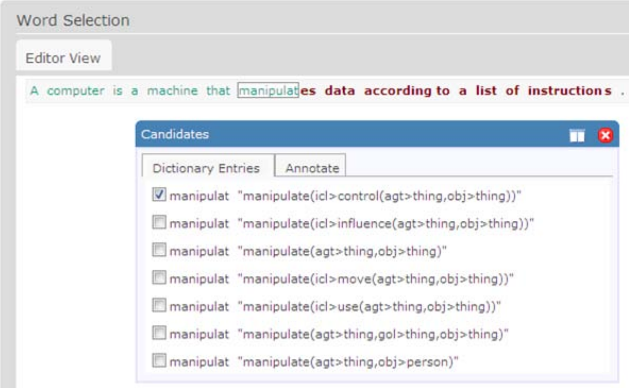
In this paper, Section 2 describes related words; Section 3 explains how our approach performs WSD; Section 4 shows some preliminary experiments and results; and finally in Section 5 we present conclusions and future work.

Contact: Francisco Tacoa, The University of Tokyo, 7-3-1 Hongo Bunkyo-ku Tokyo 113-8656, 03-5841-6774, Fax: 03-5841-8570, tacoa@mi.ci.i.u-tokyo.ac.jp

2. Related Work

2.1 Common Web Language (CWL) Platform

CWL Platform [1] is a web application that makes conversion of natural language input into a semantic description, which could be represented as CDL.nl, UNL, Resource Description Framework (RDF) [2], or as a network structure. The main goal of this application is to perform translation between natural languages through the employment of an intermediary pivot language. Under some cases, translation might present inaccuracy due to the existence of words with multiple meanings. Therefore, the application includes a module called "Word Selection" (see Figure 3) for WSD tasks. Since human intervention is required in order to use this module the process is completely manual, which makes users to have a heavy work in the case of many sentences.



A computer is a machine that manipulates data according to a list of instructions .

☑ manipulat "manipulate(icl>control(agt>thing,obj>thing))"
☐ manipulat "manipulate(icl>influence(agt>thing,obj>thing))"
☐ manipulat "manipulate(agt>thing,obj>thing)"
☐ manipulat "manipulate(icl>move(agt>thing,obj>thing))"
☐ manipulat "manipulate(icl>use(agt>thing,obj>thing))"
☐ manipulat "manipulate(agt>thing,gol>thing,obj>thing)"
☐ manipulat "manipulate(agt>thing,obj>person)"

Figure 3. "Word Selection" module in CWL Platform

CWL Platform serves as a machine translation system for a total of 6 languages: English, Japanese, French, Russian, Spanish and Arabic.

2.2 WSD and Semantic Role Labeling (SRL)

Works from [Moreda 2004] and [Moreda 2006] made combination of WSD and SRL in Question Answering and Information Retrieval systems. These works perform three main steps: (1) disambiguation of verb senses, (2) disambiguation of arguments, and (3) disambiguation of semantic roles. On the other hand, our method has been developed considering to be used in Machine Translation systems.

2.3 Selectional Preferences

Some methods implement selectional preferences as a way to constrain meanings of words. For instance, [Resnik 1993] designed a method for with verbs and semantic class of verbs' noun arguments; [Resnik 1997] exploited the verb-object and verb-subject relations; and [Agirre 2001] worked proposed class-to-class selectional preferences for WSD.

3. Our Approach

The work in this paper aims to reduce this load of work by calculating best candidates for word meanings. For this purpose, we combine analysis of syntactic relations and word-to-class selectional preferences for verb-noun relations.

The WSD method presented in this paper works with the following tools:

- Data source: an ontology of concepts and semantic relations that connect concepts pairs, known as UNL Knowledge Base (UNLKB)[3, 4]. Concepts in UNLKB are also known as Universal Words (UWs). Semantic relations are included in the CDL.nl specifications (44 in total [Uchida 2005]) and provide UWs with logical constraints.
- Syntactic parser: provides information about words, such as lemma, part of speech, and syntactic relations. All this information is necessary for the semantic analysis carried out by our method. We use Stanford Parser[5] in order to get the syntactic information.

Consider the sentence: "John eats apples". Figure 4 shows an example of UWs for verb "eat" and nouns "John" and "apple" (UNLKB contains concepts in the form of lemma):

```
eat(agt>person,obj>food)
John(iof>person)
apple(icl>fruit)
```

Figure 4. UWs in UNL Ontology

In the previous figure, "icl" is a relation that indicates a concept included in another ("apple" is included in "fruit", and "fruit" is included inside "food"); and "iof" is a relation describing that a concept is an instance of another ("John" is instance of "person"). Therefore, UW for "eat" can accept "John" as agent and "apple" as object.

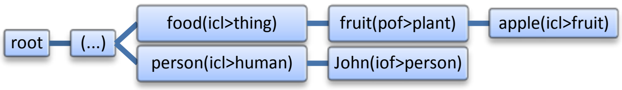


Figure 5. Word-class relation for nouns in UNL Ontology

Some other possible candidates for verb "eat" include the following UWs:

```
eat(agt>person)
eat(agt>person,obj>food)
eat(agt>thing,obj>thing)
```

Figure 6. UWs candidates for verb "eat"

[1] http://www.undl.org:8080/cwl/
[2] http://www.w3.org/RDF/

[3] http://www.undl.org/unlsys/uw/UNLKB.htm
[4] http://www.undl.org/unlexp/
[5] http://www-nlp.stanford.edu/software/lex-parser.shtml

The approach presented in this paper goes through the following 4 main steps:

- Syntactic analysis: Get information of words, such as lemma, part of speech, syntactic relations. The syntactic relations that are relevant in this case are of verb-noun and noun-noun types.
- Extraction of verb and noun candidates: Words lemmas are used to extract UWs from UNL Ontology. For each lemma, multiple results can be returned, depending on the number and type of semantic relations that the UWs have. This difference of relations is what originates the ambiguity of words that we try to solve with our approach.
- Analysis of verb-noun relations:
  - *a) Filter verb candidates:* Consider only those candidates whose semantic relations are contained in the list of syntactic relations.
  - *b) Determine best candidates for nouns:* Best candidates for nouns can be calculated by their distance to the corresponding noun class, through equation (1):

$$BC_{Noun} = min\left(dist\left(NC, N_{c_1}\right), \dots, dist\left(NC, N_{c_n}\right)\right) \quad (1)$$

where NC is the noun class and N_{c_i} is the noun candidate. Equation (1) is repeated for each noun candidate.

  - *c) Determine best candidate for verbs:* Best candidates for verbs are subject to equations (2) and (3). However, (3) will be applied only if (2) returns more than one possible candidate:

$$BC_{Verb} = max(TCR_1, TCR_2, \dots, TCR_n) \quad (2)$$

$$BC_{Verb} = min\left(\sum MD_{maxTCR\,1}, \dots, \sum MD_{maxTCR\,n}\right) \quad (3)$$

- where TCR_i means the total of connected relations for candidate i , and $MD_{maxTCR\,i}$ represents the sum of the minimum distances for the candidate i with maximum total connected relations.

- Analysis of noun-noun Relations:
  - *a) Determine best candidate for nouns:* In case that nouns without best candidates still remain, their best candidates will be calculated based on their distance to the best candidates of other already processed nouns. For this, equation (4) is used:

$$BC_{N_2} = min\left(dist\left(BC_{N_1}, N_2c_1\right), \dots, dist\left(BC_{N_1}, N_2c_n\right)\right) \quad (4)$$

where $BC_{N_1}$is the best candidate of an already processed noun, and N_2c_i represents each candidate of the noun for which the best candidate is being calculated. The noun pairs to be processed depend on the nouns connected by syntactic relations.

4. Experiments and Results

UNL Ontology is a resource still under growth, and currently is experimenting data acquisition. Therefore, it has not been possible to build a complete set of sentences for testing our method. Currently, we have a set of 160 sentences but, as consequence of missing concepts from UNL Ontology, only 33 have been successfully processed.

We used accuracy as the measure for the method evaluation, that is, whether the method could successfully determine the best candidates for word meanings. For the 33 processed sentences, our preliminary results showed an overall accuracy of 66.28%.

5. Conclusions and Future Work

This paper presented a WSD approach based on selection of best candidates for semi-automatic conversion of NL text into CDL format. The approach consists of a relations analysis method, which provides a way for best candidates' calculation.

The experiments in this work produced some preliminary results that we intend to extend as long as more data becomes available in the UNL Ontology. Results suggest that the employment of a proper correspondence of syntactic and semantic relations may contribute to a disambiguation with high precision. Moreover, it is necessary that the source of data contains the adequate concepts defined in its structure.

As future work, it has been considered to include analysis of statistical data, in order to improve the performance of the WSD approach. Most of the coming tasks will be focused on this goal.

References

[Agirre 2006] E. Agirre and P. Edmonds, Eds., Word Sense Disambiguation: Algorithms and Applications. Springer, 2006, vol. 33.

[Agirre 2001] E. Agirre and D. Martinez, Learning class-to-class selectional preferences, Proc. of the 2001 Workshop on CoNLL, vol. 7, 2001, pp. 1-8.

[Moreda 2006] P. Moreda and M. Palomar, The role of verb sense disambiguation in semantic role labeling, Advances in Natural Language Processing, vol. 4139, pp. 684-695, 2006.

[Moreda 2004] P. Moreda Pozo, M. Palomar Sanz, and A. Suárez Cueto, Assignment of semantic roles based on word sense disambiguation, IBERAMIA 2004, vol. 3315, pp. 256-265, 2004.

[Resnik 1993] P. Resnik, Selection and information: A class-based approach to lexical relationships, PhD Thesis, University of Pennsilvania, 1993.

[Resnik 1997] P. Resnik, Selectional preference and sense disambiguation, Proc. of the ACL SIGLEX Workshop, Washington, 1997, pp. 52-57.

[Uchida 2005] H. Uchida, M. Zhu, and T. Della Senta, The Universal Networking Language, 2nd ed. UNDL Foundation, 2005.

[Yokoi 2005] T. Yokoi, H. Yasuhara, H. Uchida, M. Zhu, and K. Hasida, CDL (Concept Description Language): A common language for Semantic Computing, WWW2005 (SeC2005), Makuhari, Japan, 2005.