E-055

Annotating Semantic Structure of Web Text based on CDL.nl

Yulan Yan[1], Yutaka Matsuo[1], Mitsuru Ishizuka[1], Toshio Yokoi[2]

1. Abstract

Confronting the challenges of annotating naturally occurring text into a semantically structured form to facilitate automatic information extraction, based on the Concept Description Language for Natural Language (CDL.nl) which is intended to describe the concept structure of text using a set of pre-defined semantic relations, we develop a parser to add a new layer of semantic annotation of natural language sentences as an extension of SRL (Semantic Role Labeling). The parsing task is a relation extraction process with two steps: relation detection and relation classification. We advance a hybrid approach using different methods for two steps: first, based on dependency analysis, a rule-based method is presented to detect all entity pairs between each pair for which there exists a relationship; secondly, we use a feature-based method to assign a CDL.nl relation to each detected entity pair using Support Vector Machine. We report the preliminary experimental results carried out on our manual dataset annotated with CDL.nl relations.

2. Introduction

With the dramatic increase in the amount of textual information available in digital archives and on the WWW, interest in techniques for automatically extracting information from text has been growing. Identification of information from sentences and their arrangement in a structured format to be queried and used in semantic computing applications such as web searching and information extraction are expected. Recently, much attention has been devoted to Semantic Role Labeling (SRL) of natural language text with a layer of semantic annotation having a predicate-argument structure, so-called shallow semantic parsing, which is becoming an important component in NLPs of various applications. Currently, SRL is a well-defined task with a substantial body of work and comparative evaluation [1, 2].

Although Semantic Role Labeling specifically examines predicate-argument structure, towards the goal of putting the whole sentence into a semantic structural form, Yokoi et al. (2005)[3] presented a descriptive language named Concept Description Language for Natural Language (CDL.nl), which is part of the realization of spirits of the work "semantic information processing"[4]. In fact, CDL.nl defines a set of semantic relations[1] to form the semantic structure of natural language sentences in a graphical representation. They record semantic relationships showing how each meaningful entity (nominal, verbal, adjectival, adverbial) relates semantically to another entity. It connects all meaningful entities into a unified graphical representation, not only predicate-argument related entities.

Consequently, using the CDL.nl relation set, the task of structure annotation becomes a relation-extraction process that is divisible into two steps: relation detection, which is detecting entity pairs for which each there exists a meaningful relationship; and relation classification, which is labeling of each detected entity pair with a specific relation. For CDL.nl relation extraction, the challenge we must confront is that not only the relation detection step is more difficult than a classification problem as in semantic role labeling, but also that classification of a wide variation of CDL.nl relation types is harder than that of only predicate-argument roles. In this paper, we describe a hybrid approach using two different methods for each step.

Our contributions can be summarized as the following.

- We develop a parser to add a new layer of semantic annotation of natural language sentences. Annotation of text with a deeper and wider semantic structure can expand the extent to which shallow semantic information can become useful in real semantic computing applications such as Web Search and Information Extraction.
- Our study shows an intermediate phase in the progress to semantic parsing of natural language processing from syntactic processing. It will be useful to various NLP applications such as machine translation and natural language understanding.

The remainder of this paper is organized as follows. Section 3 proposes our hybrid method for relation extraction. Section 4 reports our preliminary experimental results. We conclude our work in Section 5.

3. Hybrid Approach for Automatic Relation Extraction

We present a hybrid approach with different methods for two steps: first, based on dependency analysis, a rule-based method is advanced for relation detection; secondly, we use a feature-based method to assign a CDL.nl relation to each detected entity pair.

3.1 Rule-based Entity Pair Identification

To find a relationship between entities in the level of semantic processing, we use dependency analysis as the basis to perform our relation detection task because it shows the head-modifier relations between words in the level of surface-syntactic processing in a word-to-word way. We present a rule-based method for relation detection that is done with a simple algorithm:

Step 1: For each input sentence, generate a dependency tree that specifies the syntactic head of each word in the sentence.

Step 2: Find a headNode set from the dependency tree. Each can be a headword of a head entity to govern a relation. We select nodes which have subtrees and omit those which cannot be headNodes by creating a head stoplist.

[1] The University of Tokyo
[2] Tokyo University of Technology

[1] http://www.miv.t.u-tokyo.ac.jp/mem/yyan/CDLnl/

Step 3: For each headNode, check each of its subtrees to find those which can be tail entities related to the headNode. We create a tail stoplist containing those which cannot be root nodes of subtrees of tail entities. We continue to check the immediate grandchildren until reaching the leaf nodes if the root node of a subtree is in the tail stoplist.

Step 4: A simple post-processing is applied to correct the boundaries within which the dependency tree does not show correct relationships.

3.2 Machine Learning Method for Relation Classification

When all entity pairs have been detected, facing the challenges of labeling each pair with a specific CDL.nl relation, we describe a feature-based relation classification method that uses features to represent diverse knowledge of three levels of language processing: syntactic analysis, dependency parsing, and lexical construction.

As a benefit from the Connexor Parser[1], rich linguistic tags can be extracted as features to classify relations between entities. For each pair of entities of relation instances, we extract a syntactic feature set F_S, dependency feature set F_D, lexical feature set F_L.

Having defined all the feature sets, we develop a composite feature set to combine and leverage individual sets:

$$F_{SDL} = F_S \cup F_D \cup F_L$$

4. Experiments

Because this is the first work to extract CDL.nl relations from plain form text, currently no dataset exists for us to use for training and testing. After 46 person-days of discussion and manual annotation effort, we created a dataset[2] containing about 1700 sentences from Wikipedia documents. It was annotated with 13487 CDL.nl relations including 44 relation types. We evaluated the systems using ten-fold cross validation using this dataset. The classifier evaluation is carried out using SVM-light software [5] with our syntactic, dependency, and lexical features.

Tab. 1 Overall performance of relation extraction

TASK	Precision	Recall	*F*-value
Relation Detection(RD)	62.65	68.33	65.37
Relation Classification(RC)	86.35	87.43	86.89
RD + RC	51.62	57.94	54.60

Tab. 1 presents the preliminary result of the overall performance of our relation extraction approach by combining two steps. While the performance of the relation classification step is quite adequate, the performance of relation detection is low. Despite confronting so many obstacles, CDL.nl relations were extracted using our approach with Precision, Recall, and *F*-values that are, respectively, 51.62, 57.94, and 54.60. Data analysis reveals that aside from dependency analysis, our method of relation detection can be improved by integrating diverse information from different levels of natural language processing.

1 www.connexor.com

2 http://www.miv.t.u-tokyo.ac.jp/mem/yyan/CDLnl/

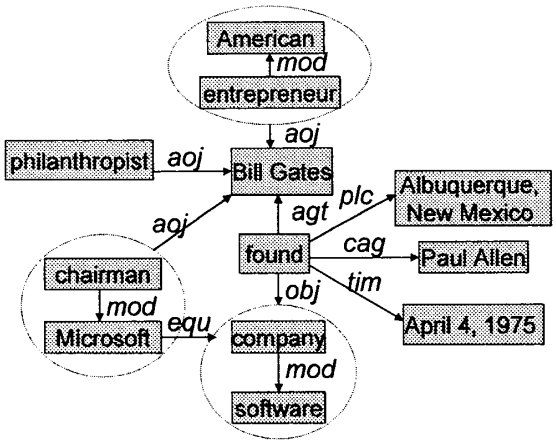

Fig. 1 The graph structure of an example sentence

Finally, Fig. 1 shows a graph describing the structure of a sentence "Bill Gates is an American entrepreneur, philanthropist and chairman of Microsoft, the software company he founded with Paul Allen in Albuquerque, New Mexico on April 4, 1975" where relations are classified and labeled by our algorithm. It shows that with CDL.nl relation set, unstructured sentence can be annotated into graph structure with not only predicate-argument relations, but also those between each pair of entities there exists a meaningful relationship, such as the *equ*(equivalent) relation between entities "Microsoft" and "the software company" shows that both refer to the same object, and *aoj*(thing with attribute) relation between "American entrepreneur" and "Gates".

5. Conclusions

In this paper, to surmount the challenges of semantic annotation of text, we created a new parser that (1) used a new set of semantic relations of CDL.nl, which has better coverage than those of SRL, to represent the semantic structure of text. In addition, (2) we proposed a hybrid relation extraction approach using two methods to detect all entity pairs between each of pair for which there exists a relationship and assign a CDL.nl relation to each detected entity pair respectively. Experiments conducted using our manual dataset revealed that our approach works better to achieve relation classification than to achieve relation detection, which can be improved by integrating diverse levels of information from natural language processing.

References

[1] K. Litkowski, "Senseval-3 task automatic labeling of semantic roles," *In Senseval-3.*

[2] X. Carreras and L. Marquez, "Introduction to the CoNLL-2005 shared task: Semantic role labeling," *In Proc. CoNLL-05.*

[3] T. Yokoi, H. Yasuhara, H. Uchida, et al. *CDL (Concept Description Language): A Common Language for Semantic Computing. In WWW2005 Workshop on the Semantic Computing Initiative (SeC2005).*

[4] M. Minsky, *Semantic Information Processing. MIT Press, Cambridge, MA.*

[5] T. Joachims, "Text Categorization with Support Vector Machine: learning with many relevant features," *In Proc. ECML-98.*