A cognitive approach for the identification of emerging entities in high-quality texts

Nivo RANDRIAMBOLOLONA , Falimanana RANDIMBINDRAINIBE

Laboratoire de recherche en sciences cognitives et applications (LRSCA)
Ecole Supérieure Polytechnique d'Antananarivo
Université d'Antananarivo, Madagascar
nivoran@gmail.com falimanana@mail.ru

ABSTRACT

This paper presents a method for identifying emerging named entities in news articles. It is based on a formal grammar that generates the language of the most frequently used named entity context schemes. Jointly used with a logistic model tree, this method brings us further by allowing the semantic classification of regular and irregular context schemes . Knowing the semantic class of a scheme affords to interpret it into knowledge and knowledge is that what is required for identification. This method represents the keystone of a project called AIDA-for-News that was motivated by the original MPII-Germany project AIDA[1].

1. Introduction

Named entity recognition cannot be dissociated from any natural language processing task in presence of proper nouns [2]. Most methods tackle this issue with probabilistic approaches. State-of-the-art NERD systems like AIDA use on one hand a CRF-based technique [3] for entity recognition and on the other hand empiric measures for disambiguation and achieve excellent precision rates on Wikipedia texts. However, this precision drops when applied on a wider variety of texts with less popular entities or emerging names. The main purpose of our work is about including emerging entities in the identification process. The basic idea is to detect and to identify them by means of a cognitive approach based on a formal grammar, which is supposed to generate the language of entity context schemes. For efficiency, we chose for the semantic classification of the produced schemes a learning based solution. The most suitable algorithm we experimented with, is a logistic model tree.

2. A context centered approach

By definition, identification consists, for a given mention m, of the selection of a single corresponding canonical entity e_i among a set of candidate entities. But what happens if no corresponding entity yet exists for a given mention? AIDA's disambiguation function combines prominence priors with context similarities and entity-entity coherence models. Unfortunately, none of these measures are intended to be applicable on emerging entities since the latter may be empirically invisible. Even worse, in case of ambiguity between an emerging entity and a popular one, the popular reference will systematically wrongfully win. To catch up this empiric deficiency, we suggest first a large cognitive exploitation of any expression that may represent an entity semantic context in a text. Later, we will combine it with a learning process for efficient semantic classification of these contexts. But first of all, for any cognitive approach to work, we need an experimentation field that offers high-quality texts in terms of linguistic, grammar and redaction styles. Furthermore, the texts should contain as much (real) emerging entities as possible. The category of texts that fills both conditions is the category of news articles.

Note : For convenience, we will use the following abbreviations : NE (named entity), NEC (named entity context), NECS (named entity context scheme).

To resume, the essence of our approach is to detect potential NEC in each sentence of a given article, to make abstraction of the underlying NECS, to associate each of them with a semantic class and finally to integrate emerging NE as new canonical entities into a knowledge base.

3. Method for context-scheme modelization

3.1 Linguistic analysis

We start with a linguistic analysis of news articles in order to detect typical morphosyntactic structures for NE-descriptions. We call context any expression that surrounds a NE-mention (including the mention itself) and that brings additional information about that NE. Three main description styles were discovered through the analysis:

- The prefix description, where the describing context precedes the NE-mention, example: *US President Joe Biden*
- The suffix description, where the describing context appears after the NE-mention. For example: *Joe Biden, the president of the USA*
- The indirect description (most of the time biographical nature), where a person is indirectly described through the description of another person. For example: *Her mother, Kimberly Schmidt, lived in Vietnam.*

3.2 Generic pattern identification

After context recognition, abstraction is made of the underlying morphosyntactic structure (vector) which we call the NE-context-scheme (NECS).
Inside a description category, schemes apparently differ from each other only at certain positions of the vectors. Such schemes can be brought together under a common generic pattern in which a variable symbol X is placed at the position that represents the branching point.
For example, let us consider the prefix-description-style to describe a person. To express the prefix-description-style, the generic pattern <X><P> is used where <X> is variable while <P> is fixed. Effectively, X could be replaced by the empty word, or by an adjective preceding the name, or by other types of parts of speech.

3.3 Grammar construction

Finally, a formalism should be used in order to integrate the generic patterns and their subexpressions into a single structure. Therefore, a formal grammar seems to be the most appropriate option. We omit deliberately the presentation of this grammar in this paper. Let us just call it the grammar G.

4. Context-scheme classification

We say a NECS is regular, if it was produced by our grammar G. Each regular NECS can be manually assigned to a semantic class from a finite set of classes. Each of these classes refers to a particular description type and each description type has an abstract semantic interpretation. We postulate that semantic classes can be predicted by means of supervised learning and that the NECS jointly with their

respective classes are best suited to serve as a training set for context-scheme classification. This will give us the possibility to learn and to classify regular and even irregular NECS.

Given the structure of the generic patterns, it seems intuitive that the classification algorithm should be based on a decision tree approach [7]. We experimented first with a classical C4.5 based decision tree on the Weka workbench. For better performances, decision trees are often combined with other learning techniques. Since our semantic classes are numbered from 1 to N, we experimented numeric prediction with a logistic model tree. We restarted the experience with the same training set but with the LMT algorithm. The obtained results tell us that the numeric prediction with LMT performs better on our training set as a classical decision tree although the issue is actually a classification problem.

5. Reification

The final goal of our work is the (semantic) identification of named entities in high-quality texts like news articles independently of their notoriety, such that even emergent entities can be identified and correctly acknowledged. To ensure that no entities get excluded, they run through the same process until their reification in a local knowledge base. The description of their reification is out of the scope of this paper.

6. Conclusion

The results of our experience showed us that the cognitive approach allows to identify more named entities than empiric approaches. The difference is clearly made by the emerging entities. As long as a context is delivered along with a name, it is possible to create a corresponding canonical entity. Beyond the identification of emerging entities, the use of semantic schemes brings another advantage. It allows in some cases (not yet systematically unfortunately) to thwart polysemy. In a news article about the hurricane *Matthew* for example, our prototype could recognize that *Matthew* is not a person thanks to the context Hurricane. Another empiric based system interpreted it as the person Matthew C. Perry, a commodore of the US Navy in the 19th century. The obtained results are encouraging because they confirm the reliability of the cognitive approach for emerging entity identification. Nevertheless, existing annotation systems are not yet prepared to treat systematically the polysemy of natural languages Tackling this problem is certainly an interesting challenge and should be assigned the first priority in the perspectives.

REFERENCES

[1]J. Hoffart, M.A. Yosef, I. Bordino, M. Spaniol, G.Weikum: AIDA: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables, Max Planck Institute for Informatics, Saarbrücken, Germany, 2011

[2]M. R. Vicente: La glose comme outil de désambiguïsation référentielle des noms propres purs, Corela – Université de poitiers, 2005

[3] J. Hoffart, Y. Altun, G. Weikum. 2014. Discovering Emerging Entities with Ambiguous Names. Seoul, South Korea : Proceedings of the 23rd International Word Wide Web Conference, WWW 2014, 2014.

[4] J. Hoffart, M.A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, G.Weikum: Robust Disambiguation of Named Entities inText, EMNLP, 2011, 784-785.