Named Entity Relation Mining Using Wikipedia

Adrian Iftene[1], Alexandra Balahur-Dobrescu[1, 2]
[1]"Al. I. Cuza" University, Faculty of Computer Science
General Berthelot Street, No. 16, Iaşi, Romania
[2]University of Alicante, Department of Software and Computing Systems
Apartado de Correos 99, E-03080 Alicante, Spain
E-mail: adiftene@infoiasi.ro, abalahur@{infoiasi.ro,dlsi.ua.es}

Abstract

Discovering relations among Named Entities (NEs) from large corpora is both a challenging, as well as useful task in the domain of Natural Language Processing, with applications in Information Retrieval (IR), Summarization (SUM), Question Answering (QA) and Textual Entailment (TE).

The work we present resulted from the attempt to solve practical issues we were confronted with while building systems for the tasks of Textual Entailment Recognition and Question Answering, respectively. The approach consists in applying grammar induced extraction patterns on a large corpus – Wikipedia – for the extraction of relations between a given Named Entity and other Named Entities. The results obtained are high in precision, determining a reliable and useful application of the built resource.

1. Introduction

Within the RTE3 competition for TE recognition (Dagan et al., 2006), the performance of the system we built owed a high percentage to the rule regarding the presence of the same NE or related NEs, both in the Text (*T*) and Hypothesis (*H*). Hence, after marking the NEs in the two text snippets using Lingpipe[1], the system verified whether all NEs in *H* were also found in *T*. If that was not the case, it sought relations between the NE from *H* without correspondence in *T* and another NE from *T* using a semi-automatically acquired collection of relations among NEs. This background knowledge was not priory available, so its acquirement became a practical issue, whose solving brought significant improvements to the system performance.

Within the task of Question Answering, we used both an acronyms database, as well as the background knowledge collection of relations between NEs to expand the NEs from the questions. With the extended list of NEs obtained, we formed a query whose probability to return correct answers was greater than that of the query including only the NEs found in the question. Such an expansion proves to be useful, not only in contest tracks, as the one at QA@CLEF, where questions are formulated given a corpus like Wikipedia, and NEs are not so often replaced by their holonyms (India by Asia etc.), but even more in real-life QA systems. When using the latter, a search-query does not always have the exact same NEs that are found in the documents from which the correct answer could be extracted.

Related work includes (Hasegawa et al. 2004) and (Weaver et al. 2006).

Our program has possibility to extract from Wikipedia snippets that contains a specified NE, or the program can extract a list with NEs related to a given NE. In first case we use a grammar that is able to identify definitions contexts, and in second case we consider relations between types of NEs. For both cases we build relations between NEs and try to evaluate the results.

2. Grammar

Under the framework of the FP6 European project LT4eL[2] (Language Technology for e-Learning), an environment for collecting and (semi)automatic exploiting language resources has been created, as the main objective of the project is to provide functionalities based on language technologies and to integrate semantic knowledge in Learning Management Systems.

In order to improve the management, distribution and retrieval of the learning material by automatically attaching metadata (such as keywords and definitions) to any text, a necessary step was the observation of those metadata in the annotated corpus. Therefore, the corpus was manually annotated to keywords, definitions of various terms and semantic concepts. Using the manual annotated documents, a grammar was created for the automatic identification of definitions in texts.

For the automatic annotation of the definitions found in the learning objects, the approach throughout the LT4eL consortium was to develop local grammars for the 9 represented languages (English, Dutch, German, Polish, Bulgarian, Maltese, Czech, Romanian, and Portuguese) to extract definition patterns.

The linguistic information from the manually annotated definitions is used as starting point in identifying possible grammar patterns that could form a definition. Previous work within this area shows that the use of local grammars which match syntactic structures of defining contexts are really useful when deep syntactic and semantic analysis is not present (Mureşan and Klavans 2002, Liu et al., 2003).

[1] Lingpipe – http://www.alias-i.com/lingpipe/

[2] LT4eL – http://www.lt4el.eu/

2.1 Categorization of Definitions

Definitions have been categorized in six types in order to reduce the search space and the complexity of rules. The types of definitions observed in our texts have been classified as follows:

- "**is_def**" - Definitions containing the verb "is".
- "**verb_def**" - Definitions containing specific verbs, different by "is". The verbs identified are "*denote*", "*show*", "*state*", "*represent*", "*define*", "*specify*", "*consist*", "*name*", and "*permit*".
- "**punct_def**" - Definitions which use punctuation signs like the dash "**-**", brackets "()", comma "**,**" etc.
- "**layout_def**" - Definitions that can be deduced by the layout: they can be included in tables when the defined term and the definition are in separate cells or when the defining term is a heading and the definition is the next sentence.
- "**pron_def**" - Anaphoric definitions, when the defining term is expressed in a precedent sentence and it is only referred in the definition, usually pronoun references.
- "**other_def**" - Other definitions, which cannot be included in any of the previous categories. In this category are constructions which do not use verbs as the introducing term, but a specific construction, such as "*i.e.*"

3. Extracting from Wikipedia NEs Related to a Specified NE

For a specified NE, we use a module to extract from Wikipedia[3] snippets with information related to it. In the snippets extracted from Wikipedia we try to identify the definition contexts. For each such context:

a) we identify the "***core***" of the definition (which is either the verb "to be" or another definition introducing verb or a punctuation mark).

b) we extract from the ***left hand*** part of the "core": all the name entities (left NEs)

c) we extract from the ***right hand*** side of the "core": all name entities (right NEs)

d) we compute the ***Cartesian product*** between left NEs and right NEs and add the resulting pairs to the existing background knowledge base.

Subsequently, we use this file with snippets and the patterns built using existing grammar in order to identify the relations between the entities. The goal in this endeavor is to identify a known relation between two NEs. If such a relation is found, we make the association and save it to an output file.

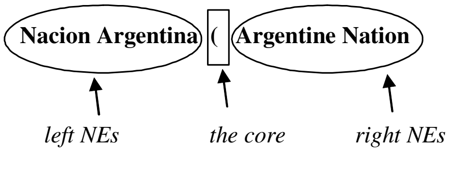

Figure 1: Example for Argentina

In Figure 1 for sentence "*Argentina, Nacion Argentina (Argentine Nation for many legal purposes), is in the world.*" we marked these values.

Eventually, only line "Argentina [is] Argentine" is added to the background knowledge.

The first level of NEs for the candidate NE is made up of all the NEs extracted for that NE (see Figure 2). Further on, all NEs related to NEs from the first level give the second level of NEs. This process is continued until no new NEs related to any of the previous levels for a candidate NE are obtained.

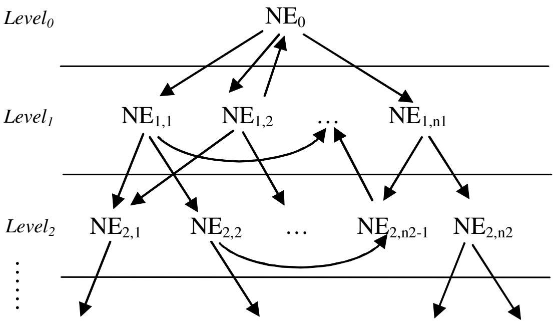

Figure 2: Levels for related NEs

Possible situations:

1) for a NE we have some related NEs on the next level
2) for a NE is possible to have some related NEs on the same level
3) for a NE is possible to have some related NEs on the previous levels

The recursive process progresses only for case 1).

A suggestive example is obtained for the NE "Paris", of type LOCATION, used as starting NE. A partial result is shown in Figure 3).

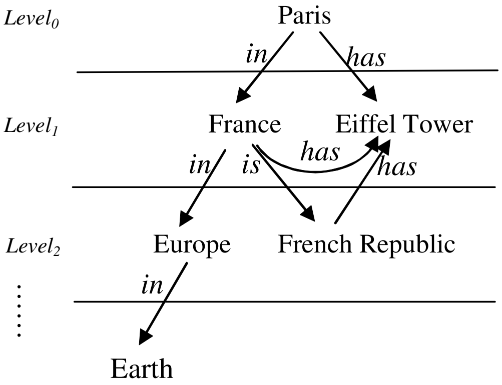

Figure 3: Example for Paris

[3] http://en.wikipedia.org/wiki/Main_Page

4. Types of Relations Between the Extracted NEs and the Initial NE – Qualitative Evaluation

The NEs extracted from Wikipedia related to a NE are clustered and classified. In order to do that we use GATE and identify the following types of NEs: country, city, ocean, sea, river, mountain, region (small and big), language, money, person name, organization, and job.
Classification depends on the initial NE type and the types of NEs related to it.
As an example, for the start entity "*Italy*" whose type is country, we extract 186 entities, from which 61 are different. The results are presented in the table below (we specified the frequency of entity appearance using the superscript notation, in the case of frequencies higher than one):

- *Money: Euro*^{2} , which is the present currency in Italy;
- *Persons*: without any relation to Italy (*Plato*^{2}), Italian artists *(Leonardo da Vinci*^{2}*, Dante Alighieri*^{3}*)*, Italian politicians (*Romano Prodi*^{2}*, Silvio Berlusconi*^{2}), and *Florence*^{5}.
- *Cities:* from Italy (*Ancona*^{2}*, Pisa*^{2}*, Naples*^{2}*, Bologna*^{2}*, Turin*^{3}*, Genoa*^{4}*, Venice*^{5}*, Milan*^{7}*, Rome*^{9}), others cities (*Constantinople*^{2}*, Marseille*^{3}). It is interesting to notice that the city with the highest frequency is *Rome*, the capital of *Italy*, and the rest of cities with a high frequency are big cities from this country.
- *Countries: Somalia*^{2}*, Greece*^{2}*, Slovenia*^{2}*, Germany*^{2}*, Austria*^{3}*, Switzerland*^{3}*, Albania*^{3}*, France*^{8}*. The countries with highest frequency are neighbors with Italy.*
- *Languages: Albanian*^{2}*, Slovenian*^{2}*, Corsican*^{2}*, Catalan*^{2}*, German*^{2}*, Friulian*^{3}*, Italian*^{3}*, Ladin*^{3}*, Greek*^{3}*, Sardinian*^{4}*. The languages with lowest frequency are spoken languages from neighbor countries, and languages with highest frequency are different Italian dialects spoken in regions from Italy.*
- *Regions: Molise*^{2}*, Dalmatia*^{2}*, Apulia*^{2}*, Emilia-Romagna*^{2}*, Tangentopoli*^{2}*, Aosta Valley*^{3}*, Veneto*^{3}*, Corsica*^{3}*, Piedmont*^{3}*, Liguria*^{3}*, Friuli-Venezia Giulia*^{3}*, Lombardy*^{3}*, Tuscany*^{4}*, Calabria*^{4}*, Sicily*^{7}*, Sardinia*^{7}. The regions appearing with the highest frequencies are well-known regions from Italy.
- *Vast Regions:* part of (*Western Europe*^{2}*, Europe*^{2}), neighbor with *North Africa*^{2}.
- *Sea: Mediterranean Sea*^{5} neighbor with it.

In order to obtain the results shown in the table below, we used over 1000 pairs under the form (*start entity, extracted entity*). The evaluation was performed manually for each of the encountered pairs. While in this first phase we focused on identifying types of relations, subsequently we focused on determining relations with high precision. Certainly, there are cases in which additional information can be extracted (in the above example, the city with the highest frequency is the country capital), but the generalization of such relations requires a higher number of testing examples consisting of entities of the same type.

Initial NE type	Type of related NEs	Relation	Precision
Country	Person Country Language Money	Person <was in> Country Country <neighbor with> Country Language <spoken in> Country Money <is currency from> Country	100 % 84 % 70 % 100 %
Organization	Country	Country <component of> Organization	70 %
Person	Person City Country Job Language	Person <know> Person Person <was in> City Person <was in> Country Person <work in> Job Person <spoke in> Language	100 % 69 % 75 % 100 % 100 %

Table 1: Types of relations between NEs

The precision score was calculated using the following formula:

$$precision = \frac{\sum_{correct\_entities} entity\_appearence\_number}{\sum_{all\_entities} entity\_appearance\_number}$$

In the example considered, 6 person names with the following frequencies were extracted:

- *Plato* – 2 appearances
- *Leonardo da Vinci* – 2 appearances
- *Dante Alighieri* - 3 appearances
- *Romano Prodi* - 2 appearances
- *Silvio Berlusconi* - 2 appearances
- *Florence* - 5 appearances.

Since the only person name found without any relation to Italy is *Plato,* the precision score is:

$$precision = \frac{2+3+2+2+5}{2+2+3+2+2+5} = \frac{14}{16} = 0.875$$

For start entity “Person”, we can deduce additional information: if we use the frequency for cities and countries, and select the city and country with highest number we can deduce where the Person was born or where the Person lives. For some cases is hard to identify the correct relation between NEs.

5. Comparison with WordNet – Quantitative Evaluation

As seen in the previous chapter, the extracted results capture well the relations between NEs. However, there remain certain questions, such as: The extracted NEs related to a NE and relations between them are correct to a high extent, but are they enough? How many NEs did we skip using Wikipedia? In order to answer these questions we extracted all NEs related to the given NE from the English WordNet and compared the results. In the following tests, we consider the “European Union” NE, of the type “Organization”.

Named Entity	**In WordNet and in Wikipedia**	**Additional in Wikipedia**	
		Correct	**Wrong**
European Union	Austria, Belgium, Denmark, Finland, France, Germany, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, United Kingdom	*Albania, Bosnia and Herzegovina*, Bulgaria, ***Croatia***, Cyprus, Czech Republic, Estonia, Greece, Hungary, Latvia, Lithuania, Malta, *Montenegro*, ***Republic of Macedonia***, Romania, *Serbia*, Slovakia, Slovenia, ***Turkey***	Andorra, Aruba, Canada, China, French Guadeloupe, Guiana, Iceland, India, Japan, Liechtenstein, Martinique, Monaco, Norway, Russia, San Marino, Switzerland, United States

Table 2: Comparison between WordNet and Wikipedia

It can be easily noticed that Wikipedia contains the new members of the EU: *Romania and Bulgaria*, three official candidates: *Croatia, Republic of Macedonia and Turkey*, and the countries that are officially recognized as potential candidates: *Albania, Bosnia and Herzegovina, Montenegro and Serbia*. Regarding the frequency of countries, the highest values are obtained for common countries from WordNet and Wikipedia and the lowest values for wrong additional values.

6. Conclusions

This article presents the methodology and results of finding relations between NEs using the Wikipedia corpus. Preliminary results indicate a good quality and quantity of the results and prove that a resource like WordNet cannot cover all continual changes in the world. The motivation for using Wikipedia is given by the necessity to build this kind of resources for different languages and the large availability of this resource in over 253 languages with around 10 millions of users. WordNet is an accurate and complex resource, but it only exists in 15 languages and the number of synsets is still very low in most languages except English. Our work is language independent and can be applied on any language with articles on Wikipedia. However, an important problem will always be the quality of this information. In order to increase the quality of obtained resources, WordNet can be very useful in identifying the good results.

Future work includes tests on more types of NEs and on a larger volume of data for improving the quality of the output, as well as the construction of a general resource that will be used in Question Answering when expanding the NEs in a given question.

7. Acknowledgements

The authors thank the members of the NLP group in Iasi for their help and support at different stages of the system development.

The work on this project is partially financed by SIR-RESDEC, PNCDI II project.

8. References

Dagan I., Glickman O., Magnini B. (2006) . The PASCAL Recognising Textual Entailment Challenge. In *Quiñonero-Candela et al.*, editors, MLCW 2005, LNAI Volume 3944. Springer-Verlag, pp. 177--190.

Mureşan S. and Klavans J. (2002). A Method for Automatically Building and Evaluating Dictionary Resources. *Proceedings of LREC 2002*.

Liu, B., Chin C. W., and Ng H. T. (2003). Mining Topic-Specific Concepts and Definitions on the Web. *Proceedings of the Twelfth International World Wide Web Conference (WWW'03)*.

Hasegawa T., Sekine S., Grisham R. (2004). Discovering Relations among NEs from Large Corpora. *Proceedings of ACL 2004 Conference.*

Weaver G., Strickland B., Crane G. (2006). Quantifying the Accuracy of Relational Statements in Wikipedia: A Methodology. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries.*