Semi-Automatic Resource Update for STRING

Catarina Vieira de Mendonça

IST – Instituto Superior Técnico
L2F – Laboratório de Sistemas de Língua Falada – INESC ID Lisboa
Rua Alves Redol 9, 1000-029 Lisboa, Portugal

Abstract. The Named Entities Recognition (NER) consists in identifying and classifying certain elements of the text according to a typology of entities, for example, names of people, organizations, places, events or dates. The identification and classification of Named Entities (NE) helps the systems of Natural Language Processing (NLP) in obtaining some meaning from text. The system that performs the task of REM is the STatistical and Rule-based Natural lanGuage processing chain (STRING) with the help of two types of resources: (i) lexicons and dictionaries.

(i) Lexicons are lists of NE that specify the interpretation of a segment associated with a node.

(ii) Dictionaries are lists of segments that must be recognized by the string.

By default, the resources are several multi-line files that are updated manually. The task of updating each of these files with new entities can be a monotonous, time-consuming and error-prone job. This work aims to automate the way in which the lexicons and dictionaries are updated. Making it faster and more systematic, thus less subject to errors. To do so, it was decided to use Wikipedia’s Infobox content to obtain and classify as many Entities as possible, and then the EM not yet present in the resources are added.

Keywords Named Entities (NE), Lexicons, LEMAS’ dictionaries, Wikipedia, Natural Language Processing (NLP), Portuguese

1 Introduction

The Named Entities Recognition (NER) [1, 2] consists of identifying certain elements of the text and classifying them according to a previously established typology of NE [3], the most usual classes being: person, place, organization, quantity, event, datum among others. The NER is a very important task for the NLP because it influences immensely the performance of other tasks such as question-answer systems, text summarization and indexing of documents. The system responsible for the REM task is the STRING [4], with the help of some external resources such as lexicons in XEROX Incremental Parsing (XIP) [5] and the dictionaries in the Lexical Morphological Analyzer (Lexman) [6, 7]. With the
constant changes and appearance of new NE, lexicons and dictionaries must be updated frequently. The purpose of this work is to automate the task of updating lexicons and dictionaries as much as possible, reducing the manual effort and possible errors involved in this process. To update lexicons and dictionaries to be automatic it is necessary to find a good source from which to obtain new Entities. After analysis, it was concluded that Wikipedia may be considered an adequate source for extracting this information, for the following reasons:

- Constant evolution and growth, with the creation of 112 new articles every day\(^1\);
- Contains a large number of NE in the categories that are the focus of this work, about 215 thousand NE;
- Partially structured and easy to extract information from dumps, which are XML files, released monthly.

2 Wikipedia’s Structure

Wikipedia has a great diversity of resources that can be used for the extraction and identification of mentioned entities. These resources are presented below.

- Articles
- Infobox
- Categories
- Internal connections
- Redirect connections
- Disambiguation pages

The Infoboxes were the resource chosen for the development of this work, they are normally in the upper right corner of the articles and show a summary of the main information. The primary advantage of this resource is the fact that the information is showed in a structured away and in pairs \(<\text{attribute, value}>\); In addition, it is possible to find a large number of entities and, also, it is relatively easy to identify the Infobox structure in Wikipedia’s dump.

Since Wikipedia is an open source, it regularly makes copies of the data, which are called dumps. These can have two formats: XML, which have the text and page metadata; and SQL Database Tables, which additionally have all the existing links on pages, between pages, categories, and redirects. In the XML file it is easier to access the Information.

In the XML, the infobox’s information is inside the structure “Info/” and is possible to find next to “Info/” the identification of EM’s model, which in the case of the figure bellow, it indicates that the model of the infobox is Biography. Within the infobox’s structure you can find pairs separated by “\(|\)” that describe the life and work of a person, with the struture, \texttt{attribute}'='\texttt{value}.

\(^1\) \url{https://stats.wikimedia.org/EN/SummaryPT.htm}
In terms of related work we have three types, two of them use Wikipedia as an external source and our system which does not use Wikipedia to help with any tasks. The two types of systems that use Wikipedia differ on the following:

- Those who use Wikipedia to identify and classify mentioned entities;
- And those that use Wikipedia for the creation and maintenance of gazetters or dictionaries, both being large lists of NE.

The systems that use Wikipedia to identify and classify mentioned entities are: Kazama and Torisawa (2007) [10], REMBRANDT (2008)[11], REMMA (2008)[12] and Mohamed and Oussalah (2014) [13] and the resources that they use are the following:

<table>
<thead>
<tr>
<th>System</th>
<th>Wikipedia’s Resources Used</th>
<th>Acesso</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kazama and Torisawa, 2007</td>
<td>Clues from the first article’s phrase</td>
<td>Online</td>
</tr>
<tr>
<td>REMBRANDT, 2008</td>
<td>Article categories</td>
<td>Online</td>
</tr>
<tr>
<td>REMMA, 2008</td>
<td>Clues from the first article’s phrase</td>
<td>Online</td>
</tr>
<tr>
<td>Mohamed and Oussalah, 2014</td>
<td>Infoboxes Attributes</td>
<td>Online</td>
</tr>
</tbody>
</table>

Table 1. Wikipedia Resources Used by Each System and Their Type of Processing

Also for these system the Wikipedia’s resources are used online, this means that they are used during the system’s execution.
The systems that Wikipedia uses for the creation and maintenance of gazetters or dictionaries are: Toral and Muñoz (2006) and Gamallo and García (2011) and the resources that they use are the following:

<table>
<thead>
<tr>
<th>System</th>
<th>Wikipedia’s Resources Used</th>
<th>Acesso</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toral and Muñoz, 2006</td>
<td>Article’s phrase</td>
<td>Offline</td>
</tr>
<tr>
<td></td>
<td>Name Hierarchy</td>
<td></td>
</tr>
<tr>
<td>Gamallo and García, 2011</td>
<td>Infoboxes</td>
<td>Offline</td>
</tr>
<tr>
<td></td>
<td>Article’s Categories</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Categories’ Hierarchy</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Wikipedia Resources Used by Each System and Their Type of Processing

In this case the Wikipedia’s resources are used offline, which means, outside the main system execution.

In both types of approach one thing is constant, the Wikipedia’s resources help improve the performance of these systems and are always beneficial.

4 Semi-Automatic Resource Updater Architecture

The module that updates semi-automatically the STRING resources is external to the processing chain, which means, it is not executed at the same time as the execution of the STRING, but rather when the STRING lexical updates are required. The purpose of this module is to get a large number of names, which represent the entities in Wikipedia’s dump, classify them, choose the NE that should be added to the resources of the STRING and finally, update the resources with the entities selected. Each of the tasks described above is executed by one of the following four steps:

- Preprocessing: identifies the infoboxes and extracts all the information in their content, with special emphasis on the infobox model for later use to classify the NE;
- NE category assignment: identifies the category of each NE and creates lists that group them by that category;
- Identification of NE candidates: identifies in the NE set obtained from the previous tasks, the entities that the STRING does not recognize yet and should be added to the lexicons;
- Lexicons Update: uses the previous list to update the resources.
The project has a modular structure to be able to use other resources in the future so as to obtain NE, as well as to classify NE in a different way from those available, thus allowing an easy change and evolution of the module. This modularity makes it possible, if external NE lists are provided, to start executing the module in the candidate NE Identification.

5 Evaluation

Two types of approaches were used to properly evaluate the result, the first one focuses on the percentage of the NE recognized and partially recognized, the second focuses on the entities that have been recognized.

![Overall results of the quantitative evaluation](image)

**Fig. 2.** STRING’s update resources module simplified architecture.

**Fig. 3.** Overall results of the quantitative evaluation
The execution of the STRING’s resources update module introduces an improvement in all categories both in total recognition and in the partial recognition of the NE. The percentage values seem at first glance low as they are around 2%-3% in the best cases. Firstly, STRING’s performance values are already quite high for MS recognition, running the 72.14% accuracy for MS identification and the 65.66% accuracy for the classification of MS (results obtained in Second HAREM (2008)). As a result, significant increases in STRING performance become more difficult to achieve. Secondly, since the external resources update module only helps the execution of the processing chain its impact is always lower than if there were changes in the chain itself. Nonetheless, the overall results were good and higher than expected. In terms of the NE that were recognized, files with a list of the NE recognized were created for the execution of the STRING with and without this module and compared. Overall, all the NE recognized before were still recognized with the module usage, plus new entities that were not recognized before, were now also recognized. Only for the models:

- *Municipio do Brasil*, where one totally recognized NE was lost;
- *Ator*, where one partially recognized NE was lost;
- *Biografia*, where five partially recognized NE were lost;
- *Futebolista*, where six partially recognized NE were lost;

These losses are not significant because they represent really small numbers in tens of thousands of NE that were to be recognized.

6 Conclusion

In conclusion, this new module has increased the wealth of STRING lexicons by continuing to recognize, in most cases, the previously recognized MS. However, although the evaluation was positive, there is still room to improve this work, both in the amount of NE added to the resources, and in their variety.
Bibliography