Application of Information Extraction in Large Semi-Structured Text

Valmir Macário Filho,[1], Ricardo B. C. Prudêncio[1], Francisco A. T. De Carvalho[1], Leandro R. Torres,[2], Laerte Rodrigues Júnior[2]

[1] Center of Informatics, Federal University of Pernambuco, Av. Prof. Luiz Freire, s/n
50740-540 Recife/PE BRAZIL
[2] Capital Login, R. da Guia, N 99 50.030-210 Recife/PE BRAZIL

Abstract Information extraction systems are used to extract only relevant text information in digital repositories. The current work proposes an automatic system to extract information in semi-structured official journals. The implemented system deployed different features sets and algorithms used in the classification of the fragments. The system was evaluated through experiments on a sample containing 22770 lines of the Pernambuco's Official Journal. The experiments performed revealed, in general, good results in terms of precision.

Keywords: semi-structured document, information extraction, text mining

1 Information Extraction on Official Journal

Official journals are documents that contain publications (e.g., acts, texts of new laws, edicts, decisions) of countries, states, cities and other institutions in the different branches of Executive, Legislative and Judiciary power. The task of finding specific information of interest in official journals is very difficult due to the great number of publications which are daily available. Although this task can be automated, it is possible to point out some difficulties with regard to this purpose: the lack of rigid models to organize the publications in the documents, no clear delimiters between different publications, the presence of abbreviated words, the presence of orthographic errors, among others.

Documents which present the above-cited characteristics are called semi-structured texts [1]. In order to manipulate such documents, an automatic system called Information Extraction (IE) system may be very suitable. IE systems are able to extract specific information of interest from a repository of textual documents. Each input of an IE system is a textual document and the output is a set of text fragments which correspond to data fields required by the user. The extracted fields can be either directly presented to the user or stored in a database for posterior access [2].

The current work develops an IE system to extract information from official journals by using ML algorithms. A publication of an official journal is a semi-structured text divided into five main fields: title, sub-title, notebook, city and process. Some difficulties to extract information from official journals can be mentioned here: (1) fields may present very similar patterns (e.g., the sentence "Edital de Intimação", may appear in the beginning of both fields subtitle and process); (2) absent fields; and (3) presence of abbreviated patterns (e.g., the word "Process" is in many publications abbreviated to "Proc.").

The architecture of the IE system which deployed the text classification approach for IE has three steps:

1. *Fragmentation*: the input text is broken into fragments which are the candidates for filling in the required data fields. In our domain, the fragments correspond to the text lines.

2. *Feature extraction*: a vector of features is created to describe each text fragment and it is used in the classification of the fragment. This task was accomplished by considering a domain vocabulary, regular expressions and text formatting features.

3. *Fragment classification*: a learned classifier associates each input fragment to a class label associated to a data field. In our system, there are ten possible class labels. In this step, we evaluated the use of three classifiers, each one representing a different family of learning algorithms: (1) the PART algorithm for inducing decision rules, (2) the Naive Bayes classifier and (3) the Support Vector Machine (SVM) classifier [3].

In our work, the performed experiments were based on a corpus of publications collected from the Judiciary segment of the Official Journal published by the State of Pernambuco, Brazil.The performance of the IE system was evaluated for 21 different scenarios (i.e., different combinations of feature sets *versus* classifiers). The same above scenarios were also applied to evaluate the usefulness of the Sliding Window (SW) approach. The experiments performed revealed, in general, good results in terms of precision, which ranged from 70.14% to 98.63% depending on the feature set and algorithm used in the classification of the fragments.

The implemented system deployed the text classification approach for IE, which revealed to be adequate for our purpose. We highlight that the application of text classifiers for IE in the domain of Official Journals is an original work. In our experiments, we evaluated different features sets and learning algorithms in the classification of the text fragments. We observed that an improvement in performance can be yielded when sequential information of the fragments is taken into account.

The IE system can be extended to other domains of application. Additionally, as future work, other approaches can be used to construct the feature sets. We also intend to use evaluate sequential learning algorithms, such as Hidden Markov Models and Conditional Random Fields to classify the fragments.

References

[1] Turmo, J., Ageno, A. and Català, N. Adaptive Information Extraction. *ACM Computing Surveys* 38(2):4 2006.

[2] Appelt, D. and Israel, D. Introduction to Information Extraction Technology. *IJCAI-99 Tutorial*, Stockholm, Sweden, 1999.

[3] Duda, R.O. and Hart, P.E. and Stork, D.G. Pattern Classification. *John Wiley & Sons*, 2001.