Why Reading Papers when EEYORE will do that for you!?

Behrang QasemiZadeh, Paul Buitelaar
Unit for Natural Language Processing, DERI, NUI Galway
firstname.lastname@deri.org

Abstract

This abstract introduces EEYORE web service. EEYORE is designed to help researchers exploring their domain's scientific publications by extracting key technical concepts and relations between them from an arbitrary set of input publications. The proposed system uses linguistic analysis and Machine learning techniques to perform its task.

1. Introduction

There is a renewed interest in developing systems for exploring scientific publications. In this abstract we introduce EEYORE. EEYORE is based on generic linguistic analysis tools and machine learning techniques and extracts key technical terms and relation between them from an arbitrary input set of documents. In the proposed scenario, a researcher uploads a set of publications into the system and provides the system with a few examples of the topic of his/her own interest. EEYORE then analyzes the input set of publications and provides facts about technical terms similar to the ones provided by the user.

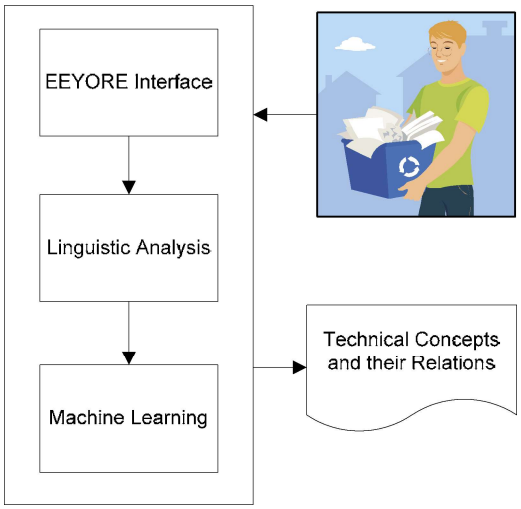

Figure 1. An Overview of the proposed research: user provides the system with a set of publications in a domain of expertise in addition to a few examples of concepts that are interesting for him/her. The system then uses the provided examples for developing a set of machine learning models that can extract concepts and facts similar to the one proposed by the user. In this way, the user may explore publications in a scientific domain in a more effective way.

2. Methodology

The proposed research is built upon two major technologies: generic human language technology [1] and machine learning techniques [2]. Detailed steps to fulfill the task are as follows:

1. **Text Extraction and Segmentation**: In this process step, scientific publications in digital format such as PDF are converted to linguistically well defined units. In other words, the input PDF files will be converted to sections, paragraphs, sentences and words.
2. **Linguistic Analysis**: The system then linguistically analyzes the input publications. The analysis includes part of speech tagging [3] i.e. the process of marking up the words in a text as corresponding to a particular part of speech, based on both its definition, as well as its context, and dependency parsing [4] i.e. a form of syntactic parsing and denotes grammatical relations between words in a sentence.
3. **Machine Learning based Classifiers:** The generated information in the previous steps of the analysis, in addition to the user provided examples/information then will be used for developing machine learning based classifiers namely a classifier for concept identification and a classifier for relation discovery.
4. **Applying Classifiers to the provided set of publications:** In this step, the system will employ the classifiers to extract new concepts and relations from the user provided publications. In addition, the system may use user feedback to refine its model and to adapt itself to user information needs.

3. References

[1] Handbook of Natural Language Processing, Second Edition, Editor(s): Nitin Indurkhya; Fred J. Damerau, Goshen, Connecticut, USA, CRC Press, 2010.

[2] Ian H. Witten and Eibe Frank. 2002. Data mining: practical machine learning tools and techniques with Java implementations. SIGMOD Rec. 31, 1 (March 2002), 76-77. DOI=10.1145/507338.507355

[3] Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-ofspeech tagger. In In EMNLP/VLC 2000, pages 63–70.

[4] Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In Proceedings of the 8th International Workshop on Parsing Technologies (IWPT), pages 149–160