Portal de Corpus: Automatic Text-Topic Identification for Generating subcorpora

Fernando Muniz and Sandra Maria Aluísio

Department of Computer Sciences, University of São Paulo
Av. Trabalhador São-Carlense, 400, 13560-970 - São Carlos/SP, Brazil
{fernando.muniz@gmail.com,sandra@icmc.usp.br}
http://www.icmc.usp.br

Abstract. In this paper, we present the Portal de Corpus, an XCES compliant corpus portal, which gives access to several Brazilian Portuguese newspaper corpora compiled in the scope of PLN-BR project. We provide several searching functions to build study corpora from a main corpus. This paper also presents our approach to allow easy access to the corpus text topics by providing content-based visual maps of the texts.

Key words: Automatic text-topic identification and Brazilian Portuguese corpora

1 Introduction

The Portal de Corpus [1][1], an XCES compliant corpus portal developed within the PLN-BR project[2], provides a tool suit for basic facilities to store and retrieve text conformant to XCES format. It was designed using open source technologies and can be easily ported to other servers. The architecture chosen was client-server, where in one side we have as interface a web site and in the other side a web server and a database. In the client side the user can access the site through a web browser. In the site, we made available an interface where users should register to have access to the main functionalities of the Portal de Corpus. A user will be able, for example, to use a Header editor to insert new texts into the database or update headers of texts already inserted. We provide several search functions to build study corpora from a main corpus based on the information present in the text headers. While the major part of the information contained in the header is based on external text information, topic is one of those that should be recovered based on internal information. Therefore after generating a subcorpus, the PEx-Corpus Tool helps the user to visually inspect the subcorpus already created to explore its content and create further subcorpora based on a selection of text topics.

[1] http://www.nilc.icmc.usp.br:8180/portal/
[2] http://www.nilc.icmc.usp.br/plnbr/

2 Header Editor, Corpus uploader and PEx-Corpus Tool

The header editor is a tool implemented using Java applets. It has a graphical interface that allows the user to create, maintain and visualize text header information that is stored in a MySQL database. To access a given database, the structure of the corpus database must follow the structure specified in this project, including the Text Typology used, which in turn follows the one used in the Lácio-Web project [2]: a four-category typology where the texts can be classified by genre, textual type, domain and medium of distribution. The header editor also has an option to insert several texts at once.

PEx-Corpus Tool employs the vector space model [3] to represent the documents as vectors in a multidimensional space, and a cosine-based distance, to determine the dissimilarities amongst the documents as the distance between the vectors that represent them. In the vector space model, the terms that occur in the collection are the space dimensions, and the frequencies of these terms in each document are the coordinates. The process used on PEx-Corpus Tool involves three main steps: (i) removing stopwords, i.e., non-informative words such as articles, prepositions and such, plus any words known to lack relevance to context (the stopword list can be defined by the user); (ii) frequency counting, so as to remove terms that occur too sparsely or too often and hence have little differential capability (Luhns cut-off); and (iii) weighting the terms according to the term-frequency-inverse-document-frequency (tfidf) measure [3]. In order to enable visual identification of the main topics discussed in the documents of the collection, a projection area can be selected - delimiting a region with the mouse - and a label is generated that is representative of the documents within this area.

In the PROPOR demonstration section we will run all steps of the corpus processing - building a corpus, header editing, building a study corpora from a main corpus and creating further subcorpora based on a selection of text topics. We outline a script of our demonstration at
http://nilc.icmc.usp.br/Portal/demo/script.htm

References

1. Muniz, M.; Paulovich, F. V.; Minghim, R.; Infante, K.; Muniz, F.; Vieira, R.; Aluísio, S. Taming the tiger topic: an XCES compliant corpus Portal to generate subcorpus based on automatic text topic identification. In: Corpus Linguistics 2007 conference, 27-30 July 2007, University of Birmingham. Proceedings of the Corpus Linguistics 2007 conference, 2007
2. Aluísio, S. M. G. Pinheiro, M. Finger, M. G. V. Nunes and S.E. Tagnin: The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In: Corpus Linguistics 2003: Proceedings of Corpus Linguistics 2003. Lancaster: 2003. v. 16, pp. 1421.
3. Salton G. (1991): Developments in automatic text retrieval. Science, 253:97480.