Individual paper proposal
Conference: TEI Conference 2014, Northwestern University, October 22-24, 2014
Title: Assisted Construction of Taxonomies for LdoD
Authors:
António Rito Silva (INESC-ID, Department of Computer Science and Engineering, IST, Universidade de Lisboa), rito.silva@tecnico.ulisboa.pt
Manuel Portela (CLP, University of Coimbra), mportela@fl.uc.pt
Keywords: Fernando Pessoa, Book of Disquiet, digital archive, taxonomies, tei-encoding, web 2.0

Assisted Construction of Taxonomies for LdoD

Context

Fernando Pessoa's *Book of Disquiet* (*Livro do Desassossego – LdoD*) is an unfinished book project. Pessoa wrote more than five hundred texts meant for this work between 1913 and 1935, the year of his death. The first edition of this book was published only in 1982, and another three major versions have been published since then (1990-91, 1998, 2010). As it exists today, *LdoD* may be characterized as (1) a set of autograph (manuscript and typescript) fragments, (2) mostly unpublished at the time of Pessoa's death, which have been (3) transcribed, selected, and organized into four different editions, implying (4) various interpretations of what constitutes this book. Editions show four major types of variation: variation in readings of particular passages, in selection of fragments, in their ordering, and also in heteronym attribution.
In the TEI 2013 conference we have presented our work on textual encoding and social editing in a Web 2.0 environment for an LdoD archive where experts and lay users can interact in the construction of virtual editions of LdoD (Silva and Portela, 2013). The full virtual model for the *LdoD Archive* is detailed in Portela and Silva, 2014.

Goals

Extracting and creating additional information based on the transcribed and encoded set of fragments (500 to 700, depending on the editors) in a Web 2.0 environment is our main goal for the current stage of the project. In the *LdoD Archive* this production of information occurs in the context of a virtual edition, where the collaborative editors (i.e. the virtual edition participants) define their own edition by choosing and ordering the fragments. Besides, they are also able to annotate parts of the fragments with comments and tags. However, the system is not yet taking advantage of the existing powerful text analysis tools to extract semantic information from the set of fragments that belong to a virtual edition. Therefore, the current goal of the project is to integrate text analysis tools with Web 2.0 technologies to enhance the capture and production of information. The integration of text analysis tools should assist a human-based process, where virtual editors can experiment with the results of text analysis, including the

possibility of redefining parameters and undoing the results of previous experiments to try out new ideas for generating taxonomies.

Problem

There is an ongoing debate in Digital Humanities about what could be the contribution of text analysis tools for humanistic methods of inquiry (Moretti 2013; Drucker 2012; Witmore 2012). Currently, most work has been done by explicitly encoding texts, what can be called 'smart data', instead of taking advantage of text mining tools to extract information, what is called 'big data' (Schöch, 2013). While the encoding of text is expensive, both in terms of time and human resources, the use of text mining approaches promises to enhance the capture of information while saving time and resources. However, there is an issue in terms of the quality of extracted data when using big data approaches (Feldman and Sanger, 2013). This problem is particularly relevant when considering *LdoD*, where the number of texts may not be significant when compared with the dimension of other corpora where these approaches are usually applied. On the other hand, there exists work on the assisted construction of ontologies, in which modelers are assisted by tools in the process of defining ontologies (Wache et al, 2001; Fortuna et al, 2006).

Solution

The solution we propose for the identified challenges is based on the assisted construction of taxonomies by integrating the algorithmic capture of knowledge, using a topic modelling strategy, with their subsequent treatment by editors. Machine text analysis and human text analysis are placed in a feedback circuit which enables the incremental modelling of taxonomies based on an automatic generation of topics.

Automatic Generation of Topics

In a first step the system generates a topic model for a selected corpus. We follow the probabilistic approach to topic modelling (Blei et al, 2003) and use its implementation in Mallet (McCallum, 2002).

The *LdoD* collaborative system supports the concept of a virtual edition. A virtual edition is the result of the social construction of an *LdoD* edition by one or more editors. These editors select the set of fragments that form the edition. This set of fragments constitutes the edition corpus and editors can generate a topic model for this corpus. The topic model generation is parameterized in terms of the number of topics to be generated, the number of words to be visualized for each topic, the percentage threshold necessary to consider a topic relevant for a fragment and the number of iterations the algorithm should execute to generate the topic model. The result of the generation is a set of topics and the association of topic with fragments, weighted by a probability. Only associations above the editors' defined threshold are considered. Additionally, editors visualize the correlated words that are relevant for the topic. The number of words for visualization is also defined by the editors when they generate the topic model.

Therefore, through this first step the editors can experiment with generation of several topic models for the corpus by choosing different parameters or changing the fragments belonging to the corpus. Additionally, the system provides an interface where they can navigate between topics and fragments. For each fragment the systems presents the associated topics with their percentage, and for each topic the set of associated fragments and words.

Incremental Modelling of Taxonomies

In the second step the editors transform topics into the categories of a taxonomy by applying their knowledge about *LdoD*. When this step starts it is considered a category for each topic generated in the previous step. These categories are incrementally changed through a set of operations available in the system interface. These operations are:

- Update the category name – when the incremental modelling step starts the category name is equal to the topic name, which is the concatenation of the topic correlated words;
- Merge categories – two categories become a single category and it also merges their associations with fragments;
- Extract a category – by selecting a subset of the fragments associated with a category it creates a new category with the associated fragments;
- Delete a category – all its fragments are dissociated;
- Associate a fragment with a category;
- Dissociate a fragment from a category.

To enhance the exploratory aspects of the incremental modelling of the taxonomy, the system also provides an undo operation for each of the above operations. Therefore the editors can, for instance, recover the categories that were previously merged.

TEI Representation of Taxonomies

Once the taxonomies are completely defined, the system allows the generation of a TEI encoded corpus for the virtual edition, which contains the defined taxonomies and their associations with each of the fragments. To represent the taxonomies we use the <taxonomy> and <category> TEI elements. Annotations are used in each of the fragments headers to encode the set of categories associated with it.

The *LdoD* system allows the definition of public and private taxonomies. The association of categories of public taxonomies to fragments is visible in the public interface of the *LdoD Archive*, but the private categories are only visible to the virtual edition editors. This kind of information cannot be encoded in the generated TEI.

Conclusions and Future Work

In this paper we describe the recent work we have done on the *LdoD Archive: A Collaborative Archive of the Book of Disquiet*. The system supports the incremental construction of taxonomies on top of a generated topic model. The system is online and highly interactive to allow both expert and non-expert users to experiment with and explore the construction of taxonomies for *LdoD* editions.

In the future we intend to integrate adhoc – i.e., where there isn't a generation step and editors can freely tag parts of fragments using an open vocabulary – with generated taxonomies. Therefore, we envision supporting the use of a restricted vocabulary to tag sentences and words in fragments. This way, it would be possible to split editors in terms of two different roles: a role to define a taxonomy, and a role to apply the categories of the taxonomy to chosen parts of the fragment.
During the conference we intend to make a more detailed presentation of taxonomy generation in the *LdoD* system and do a demo of the prototype under development.

Funding

"No Problem Has a Solution: A Digital Archive of the *Book of Disquiet*", research project of the Centre for Portuguese Literature at the University of Coimbra, funded by FCT (Foundation for Science and Technology). Reference: PTDC/CLE-LLI/118713/2010. Co-funded by FEDER (European Regional Development Fund), through Axis 1 of the Operational Competitiveness Program (POFC) of the National Strategic Framework (QREN). COMPETE: FCOMP-01-0124-FEDER-019715. This work was also supported by national funds through FCT – Foundation for Science and Technology, under project PEst-OE/EEI/LA0021/2013.

Bibliography

Blei, D., Ng, A., Jordan, M., and Lafferty, J. (2003). "Latent Dirichlet allocation". *Journal of Machine Learning Research* 3: 993–1022.

Drucker, Johanna (2012). 'Humanistic Theory and Digital Scholarship', in Mathew K. Gold, ed., *Debates in Digital Humanities*, Minneapolis: University of Minnesota, pp. 85-95. Also at: http://dhdebates.gc.cuny.edu/debates/text/34

Feldman, R., and Sanger, J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press. ISBN 978-0-521-83657-9

Fortuna, B., Mladenic, D., and Grobelnik, M. (2006). 'Semi-automatic Construction of Topic Ontologies', *Semantics, Web and Mining Lecture Notes in Computer Science* Volume 4289, 2006, pp 121-13.

McCallum, Andrew Kachites (2002). ' 'MALLET: A Machine Learning for Language Toolkit'. http://mallet.cs.umass.edu. 2002.

Moretti, Franco (2013). *Distant Reading*, London: Verso.

Portela, Manuel and Silva, António Rito (2014). 'A Model for a Virtual *LdoD*', *Literary and Linguistic Computing*, Volume 29(2) (forthcoming) DOI: 10.1093/llc/fqu004

Schöch, Christof (2013). 'Big? Smart? Clean? Messy? Data in the Humanities', in *Journal of Digital Humanities,* 2:3. 2013. http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/

Silva, António Rito and Manuel Portela (2013). 'TEI4LdoD: Textual Encoding and Social Editing in Web 2.0 Environments', Fabio Ciotti and Arianna Ciula, eds., *The Linked TEI: Text Encoding in the Web: TEI Conference and Members Meeting 2013, Book of Abstracts*, DIGILAB, Università La Sapienza, Roma, pp. 119-126. ISBN 978-88-6507-542-5.

Wache, H., Visser, U., and Scholz, T. (2001). Ontology Construction-An Iterative and Dynamic Task. FLAIRS Conference, 445-449

Witmore, Michael (2012). 'Text: A Massively Addressable Object', in Mathew K. Gold, ed., *Debates in Digital Humanities*, Minneapolis: University of Minnesota, 324-327. Also at http://dhdebates.gc.cuny.edu/debates/text/28