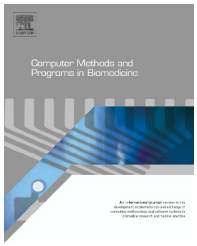
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE XXX (2014) XXX–XXX

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Marky: A tool supporting annotation consistency in multi-user and iterative document annotation projects

Martín Pérez-Pérez[a], *Daniel Glez-Peña*[a], *Florentino Fdez-Riverola*[a], *Anália Lourenço*[a,b,*]

[a] *ESEI – Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, Universidad de Vigo, 32004 Ourense, Spain*[1]

[b] *Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal*

ARTICLE INFO

Article history:
Received 24 July 2014
Received in revised form
24 October 2014
Accepted 18 November 2014

Keywords:
Document annotation
Collaborative annotation
Iterative annotation
Inter-annotator agreement
Tracking system

ABSTRACT

Background and objectives: Document annotation is a key task in the development of Text Mining methods and applications. High quality annotated corpora are invaluable, but their preparation requires a considerable amount of resources and time. Although the existing annotation tools offer good user interaction interfaces to domain experts, project management and quality control abilities are still limited. Therefore, the current work introduces Marky, a new Web-based document annotation tool equipped to manage multi-user and iterative projects, and to evaluate annotation quality throughout the project life cycle.

Methods: At the core, Marky is a Web application based on the open source CakePHP framework. User interface relies on HTML5 and CSS3 technologies. Rangy library assists in browser-independent implementation of common DOM range and selection tasks, and Ajax and JQuery technologies are used to enhance user–system interaction.

Results: Marky grants solid management of inter- and intra-annotator work. Most notably, its annotation tracking system supports systematic and on-demand agreement analysis and annotation amendment. Each annotator may work over documents as usual, but all the annotations made are saved by the tracking system and may be further compared. So, the project administrator is able to evaluate annotation consistency among annotators and across rounds of annotation, while annotators are able to reject or amend subsets of annotations made in previous rounds. As a side effect, the tracking system minimises resource and time consumption.

Conclusions: Marky is a novel environment for managing multi-user and iterative document annotation projects. Compared to other tools, Marky offers a similar visually intuitive annotation experience while providing unique means to minimise annotation effort and enforce annotation quality, and therefore corpus consistency. Marky is freely available for non-commercial use at http://sing.ei.uvigo.es/marky.

* *Corresponding author at*: ESEI – Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, Universidad de Vigo, 32004 Ourense, Spain. Tel.: +34 988 387013; fax: +34 988 387001.
E-mail address: analia@uvigo.es (A. Lourenço).

[1] http://sing.ei.uvigo.es/.

http://dx.doi.org/10.1016/j.cmpb.2014.11.005
0169-2607/

1. Introduction

Text Mining (TM) has a wide range of applications that require differentiated processing of documents of various nature [1,2]. Ultimately, the goal is to learn how to recognise and contextualise information of interest [3]. Therefore, the annotation of documents by domain experts is invaluable to provide for a ground truth against which to train and evaluate TM methods and algorithms.

Depending on the application area, the creation of such semantically annotated corpora is a resource and time consuming activity [4,5]. Usually, multiple domain experts should review the documents and manual annotations should be compared. The initial set of annotation guidelines often fails to anticipate some of the semantics issues and it is highly unlikely that multiple annotators completely agree on the annotation of a document [6–8]. So, annotation consistency needs to be monitored through the multiple rounds of the project in order to identify relevant differences in annotation patterns and make opportune amendments to the annotation guidelines and schema, and thus, guarantee the quality of the generated corpus. This entails the active monitoring and reinforcement of inter-annotator consistency (i.e. two annotators should annotate the same text fragment equally) and intra-annotator consistency (i.e. if an annotator should annotate multiple occurrences of same text fragment similarly, within the same context).

Annotation tools are expected to manage such multi-user and iterative annotation projects actively and efficiently. So far, most of the document annotation tools available can be differentiated in terms of the task-specific specialisation of the interface, i.e. the modularity and configurability of the annotation environment [9], whereas providing similar limited quality monitoring and control abilities. Just as means of comparison, Table 1 summarises the main characteristics of some of the most recent annotation tools, and the tool presented in this work. The U-Compare of the Apache Unstructured Information Management Architecture (UIMA) [10] and the Teamware of the General Architecture for Text Engineering (GATE) [11] are two meaningful examples of open-source framework-integrated document annotation tools. Argo stands as a workbench for building and running text analysis solutions [12], while MyMiner [13], EGAS [14], PubTator [15], BioNotate [16] and BioAnnote [17] are examples of free independent annotation tools, all of which are commonly used in biomedical applications. Finally, A.nnotate (http://a.nnotate.com/), MMAX2 (http://mmax2.sourceforge.net/) and annotateit/Annotator (http://annotateit.org/) are experienced broad-application systems.

Seeking to overcome some of the current limitations, this paper presents Marky, a free Web-based document annotation tool, which implements the main steps of the annotation project life cycle, with particular emphasis on annotation quality assessment. Notably, its novelty lays on (i) the annotation tracking system, which ensures that all actions occurring in the annotation project are recorded and may be amended and (ii) the annotation quality evaluation tool, which monitors inter-annotator agreement (IAA) and intra-annotator patterns.

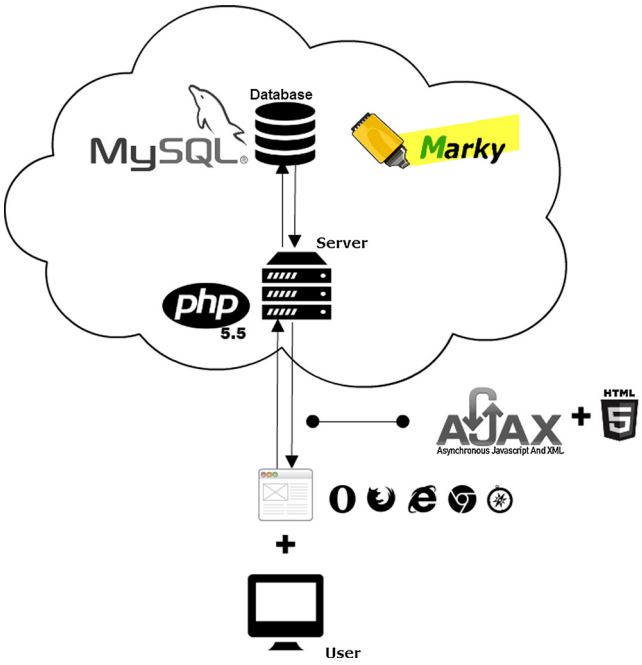


Fig. 1 – Main technologies supporting the architecture of Marky.

In the next sections, the architecture of Marky is described, in terms of the main design requirements and the implemented annotation life cycle. A case study taken from the literature is used to exemplify the abilities of Marky in practical terms. The final discussion stresses the importance of enforcing annotation consistency and assisting the work of annotators, and presents near future developments.

2. Methods

Marky is a general purpose Web-based application for document annotation. This section describes the requirements that motivated some of the aspects of its design, the overall system architecture, and the annotation life cycle implemented by the tool.

2.1. *Requirements*

From the start, Marky was designed to support general purpose annotation, i.e. domain specifications are considered only in project configuration and do not affect the behaviour of the software. This choice has allowed us to concentrate more on the infrastructure than on the application itself.

Marky is built on top of open technologies and standards to grant extensibility and interoperability with other systems. Internally, project configuration and management is kept simple, but flexible. A project may have associated multiple annotators and include several rounds of annotation. The corpus to be annotated may be loaded through a Web-accessible bibliographic service, from a local folder, or by copying contents manually. Marky handles structured documents,

Please cite this article in press as: M. Pérez-Pérez, et al., Marky: A tool supporting annotation consistency in multi-user and iterative document annotation projects, Comput. Methods Programs Biomed. (2014), http://dx.doi.org/10.1016/j.cmpb.2014.11.005