TagCurate: crowdsourcing the verification of biomedical annotations to mobile users

Bahar Sateli, Sebastien Luong, René Witte✉

Concordia University,Montréal , Canada

Received 7 August 2013; **Accepted** 12 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

We present TagCurate, a distributed system that allows for disseminating biomedical annotations to users on Android-enabled devices for further verification. A web-based interface provides Task Managers with the ability to supervise the crowdsourcing process, as well as viewing the results gathered from the TagCurate Android app installed on the crowd's devices. We believe that the results of this research is beneficial to both curators and the NLP development communities. The efforts of expert curators will be efficiently allocated to resolving controversial annotations, while NLP pipeline developers can further train their algorithms from gold standard corpora solicited from a large group of contributors.

Motivation and Objectives

Robust, automatic approaches are being developed by industrial and academic communities specifically targeting the well-known problem of information overload, caused by the overabundance of available scholarly publications. State-of-the-art approaches, in particular from the computational linguistics and Natural Language Processing (NLP) domains, aim at aiding the laborious task of manually extracting structured knowledge from the unstructured free-style text found in scientific publications. However, the inherent complexity of natural languages used by researchers in communicating their findings makes the knowledge extraction an intricate task in need of human verification to produce effective results. Based on the division of labor principle, we are proposing a novel system to *crowdsource* the verification of automatically extracted annotations from biomedical literature to mobile users. The hypothesis behind our research is that providing a synchronised, distributed system for human verification of annotations generated during the literature curation process helps to decrease the time needed to accomplish the task, while improving the curators' productivity by providing an ubiquitous environment available both in the web and mobile context.

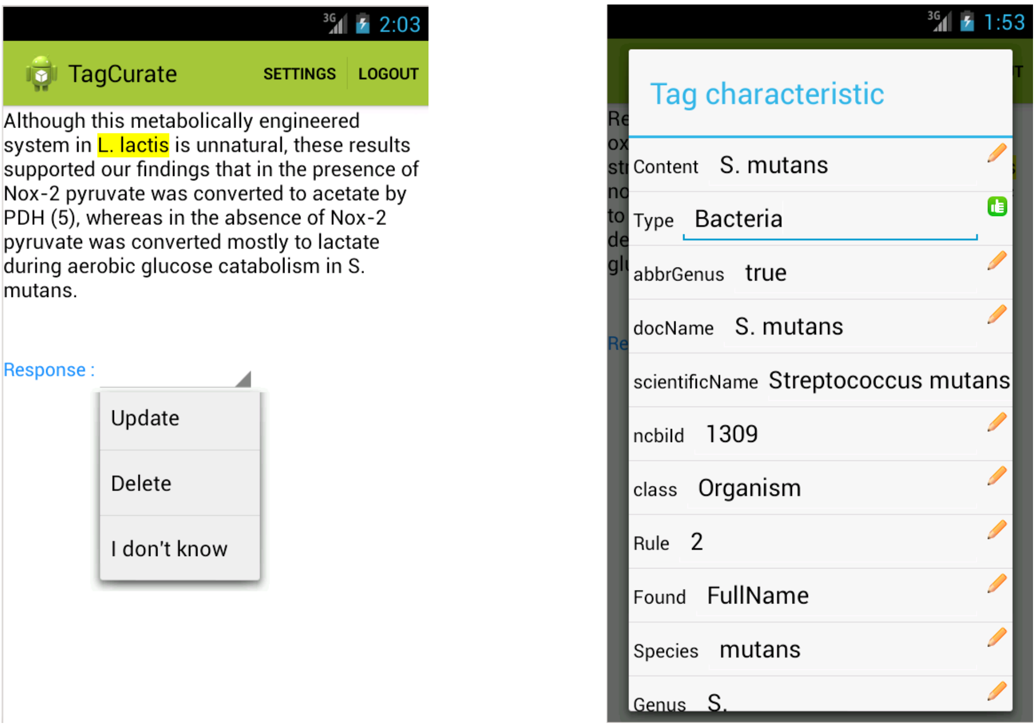
We present TagCurate, a distributed system that allows for disseminating biomedical annotations to users on Android-enabled devices for further verification. A web-based interface provides *Task Managers* with the ability to supervise the crowdsourcing process, as well as viewing the results gathered from the TagCurate Android app installed on the crowd's devices. We believe that the results of this research is beneficial to both curators and the NLP development communities. The efforts of expert curators will be efficiently allocated to resolving controversial annotations, while NLP pipeline developers can further train their algorithms from gold standard corpora solicited from a large group of contributors.

Methods

The TagCurate system is composed of a server-side component responsible for distributing and managing annotations and an Android app through which users verify the annotations assigned to them. Conforming to a client/server model, both components communicate over the HTTP protocol through a message passing mechanism.

As an extension to the Semantic Assistants framework (Witte and Gitzinger, 2008), the server-side component is implemented using the J2EE Servlet technology and provides a RESTful endpoint to interact with the TagCurate Android app. Featuring a web-based user interface, the TagCurate system allows so-called Task Managers to define a verification task by uploading annotated documents, provided that they have been annotated by either NLP pipelines or human annotators based on the General Architecture for Text Engineering (GATE) framework[1] (Cunningham *et al.*, 2011). The TagCurate system then generates an internal representation of the existing annotations, in form of an XML document, that are ultimately distributed to the TagCurate Android apps installed on the crowd's devices.

[1] http://gate.ac.uk/



Although this metabolically engineered system in L. lactis is unnatural, these results supported our findings that in the presence of Nox-2 pyruvate was converted to acetate by PDH (5), whereas in the absence of Nox-2 pyruvate was converted mostly to lactate during aerobic glucose catabolism in S. mutans.

Tag characteristic

Content S. mutans
Type Bacteria
abbrGenus true
docName S. mutans
scientificName Streptococcus mutans
ncbiId 1309
class Organism
Rule 2
Found FullName
Species mutans
Genus S.

(a) (b)

Figure 1. TagCurate Android App annotation verification (a) and modification (b) activities.

The TagCurate Android app, designed based on the Android 4.3 Jelly Bean API, allows users to authenticate themselves on the server-side component and *pull* annotations that are assigned to them for verification. Through an interactive interface, users can view each annotation in the context of the sentence that they appear, in order to determine whether the annotation is correctly tagged. A long-click on each annotation allows users to view features, i.e., further information provided in the annotation representation, like the scientific name of an organism (Naderi *et al.*, 2011), where applicable. Based on the available information, users can then decide whether the annotation is entirely correct or should be removed from the document in case of a false positive (Figure 1-a). In addition, if an annotation is partially correct, e.g., if only some features are wrong or the character offset (span) of the annotation in the text needs to be changed, users can directly edit the annotation features (Figure 1-b) and submit an updated representation to the server-side, where it is made persistent.

The gathered feedback from the crowd is aggregated on the TagCurate server-side component that can provide overview reports of the crowdsourcing progress, as well as reporting annotations with high percentage of disagreement between users, once a specified threshold is passed.

Results and Discussion

TagCurate is the first open source project that targets the distributed verification of (biomedical) annotations to mobile users. Rather than relying on expert annotators only, it is now possible to distribute document annotations, whether they are manually created or computed by an NLP pipeline, to a large user base – for example, students in a university setting. TagCurate will soon be available both as open source software and on the Android "Google Play" market for direct installation.

In future research, we will apply our mobile crowdsourcing platform to a large-scale annotation task. In particular, we will investigate incentives for mobile users to participate in verification tasks and analyse the quality differences that can be obtained from mobile users with varying backgrounds vis-à-vis expert curators. In addition to the verification task, we are also planning to provide mobile users with the ability to automatically annotate the documents using our Semantic Assistants Android Open Intents (Sateli *et al.*, 2013) that allows for remote execution of NLP pipelines on provided content, thereby enabling users to perform the complete literature curation task entirely through a mobile interface.

References

Cunningham H, Maynard D, et al. (2011) *Text Processing with GATE (Version 6)*, University of Sheffield, Department of Computer Science. 15 April 2011. ISBN 0956599311.

Sateli B, Cook G, and Witte R (2013) Smarter Mobile Apps through Integrated Natural Language Processing Services. In *10th International Conference on Mobile Web Information Systems (MobiWIS 2013)*, Paphos, Cyprus, Springer Lecture Notes on Computer Science LNCS 8093, pp. 187--202. August 26--28, 2013, doi:10.1007/978-3-642-40276-0_15

Naderi, N, Kappler T, Baker CJO, and Witte, R (2011) OrganismTagger: Detection, normalization, and grounding of organism entities in biomedical documents. *Bioinformatics* **27**(19), 2721-2729. doi:10.1093/bioinformatics/btr452

Witte R and Gitzinger T (2008) Semantic Assistants - User-Centric Natural Language Processing Services for Desktop Clients, In *Asian Semantic Web Conference (ASWC 2008), Springer Lecture Notes on Computer Sciences LNCS* **5367**, 360–374. doi:10.1007/978-3-540-89704-0_25