Annotating biomedical ontology terms in electronic health records using crowd-sourcing

Andre Lamurias [1,2]*, Vasco Pedro [3], Luka Clarke [2] and Francisco M. Couto [2]

[1]BioISI: Biosystems & Integrative Sciences Institute, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2]LaSIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749-016, Lisboa, Portugal
[3]Unbabel, 360 3rd Street, Suite 700, San Francisco, CA 94107-1213, USA

ABSTRACT

Electronic health records have been adopted by many institutions and constitute an important source of biomedical information. Text mining methods can be applied to this type of information to automatically extract useful knowledge. We propose a crowd-sourcing pipeline to improve the precision of extraction and normalization of biomedical terms. Although crowd-sourcing has been applied in other fields, it has not been applied yet to the annotation of health records. We expect this pipeline to improve the precision of supervised machine learning classifiers, by letting the users suggest the boundaries of the terms, as well as the respective ontology concept. We intend to apply this pipeline to the recognition and normalization of disorder mentions (i.e., references to a disease or other health related conditions in a text) in electronic health records, as well as drug, gene and protein mentions.

1 INTRODUCTION

Electronic health records (EHRs) are a source of information relevant to various research areas of biomedicine. These records contain details on diseases, symptoms, drugs and mutations, as well as relations between these terms. As more institutions adopt this type of system, there is an increasing need for methods that automatically extract information from textual data. This information may be matched to existing ontologies, with the objective of either validating the information extracted or expand the ontology with new information.

Text mining methods have been proposed to automatically extract useful information from unstructured text, such as EHR. Named Entity Recognition (NER) is a text mining task which aims at identifying the segments of text that refer to an entity or term of interest. Another task is normalization, which consists of assigning an ontology concept identifier to the recognized term. Finally, the relations described between the identified terms can be extracted, which is known as Relation Extraction.

The results of these tasks should be as accurate as possible so that minimal human intervention is required to use the results for other applications. To evaluate fairly the state-of-the-art of text mining systems, community challenges have been organized, where the competing systems are evaluated on the same gold standard. The task 14 of SemEval 2015 consisted in the NER of disorder mentions from EHR, as well as the normalization to the SNOMED-CT subset of UMLS (Campbell *et al.*, 1998). The best F-measure obtained for this task was of 75.5%. The CHEMDNER task of BioCreative IV consisted in the recognition of chemical entities in the titles and abstracts of PubMed articles. For this task, the best F-measure was of 87.39%. The difference between the results of the two tasks could be due to the fact that EHR may contain more noise than scientific articles. These results show that there is a need to improve the state-of-the-art, to satisfy the user expectations on automated extraction of biomedical information from unstructured text.

In this paper we propose a pipeline to improve the extraction and normalization of biomedical ontology terms in EHR by crowd-sourcing the validation of the results obtained with machine learning algorithms. This approach has been applied to other types of tasks, with promising results. The crowd would be used to validate the boundaries of the term, as well as the ontology concept associated.

2 NORMALIZATION OF BIOMEDICAL TERMS TO ONTOLOGIES

The results produced by NER methods may be normalized to unique identifiers from ontologies. The advantage of this approach is that the structure of the reference ontology may be used to validate the information extracted from the text. We have explored semantic similarity between chemical entities matched to ChEBI concepts, which improved the precision of our system (Lamurias *et al.*, 2015).

The normalization of entities is a challenge due to the ambiguity and variability of the terminology. The same label may refer to different concepts, depending on the context, while one concept may be mentioned with different names, due to spelling variants, abbreviations and capitalization. While the ontology may provide a set of synonyms for each concept, it is usually incomplete, requiring a method more advanced than string matching to correctly normalize an entity.

3 CROWD-SOURCING IN ANNOTATION TASKS

Text processing tasks are suitable candidates for crowd-sourcing since they cannot be solved computationally, and can be broken down into smaller micro-tasks (Good and Su, 2013). For example, it has been applied to machine translation (Ambati and Vogel, 2010), recognition of names in historical records (Sukharev *et al.*, 2014), question-answering (Mrozinski *et al.*, 2008) and ontology alignment (Sarasua *et al.*, 2012). Crowd-sourcing micro-tasks are usually defined by the large volume of tasks to be performed, as well as the simplicity of each individual task. The participants may be motivated by monetary rewards (e.g. Amazon Mechanical Turk), games with purpose (Von Ahn and Dabbish, 2008), or simply the satisfaction of having contributed to a larger project (Jansen *et al.*, 2014).

*To whom correspondence should be addressed: alamurias@lasige.di.fc.ul.pt

Computational methods to map a term to an ontology concept, usually based on string similarity, are able to find one or more matches for each term. However, a machine is not able to identify the most correct term from a list of matches with the accuracy of a human annotator. By letting a large number of participants evaluate the ontology concepts matched to the terms recognized in a given text, a new dataset can be generated with these corrections. This dataset would be used to train a classifier able to determine the correct concept corresponding to a recognized biomedical concept, such as disorder, chemical, protein or gene, with high precision. This classifier can be trained with a supervised machine learning algorithm, or with reinforcement learning. Likewise, a golden dataset could be generated to evaluate and tune the classifier.

4 PIPELINE

The pipeline is composed by two modules: one for NER of disorder, chemical, gene and protein mentions, and another for normalization to SNOMED-CT, ChEBI and Gene Ontology concepts, respectively.

The NER modules starts with classifiers trained with existing annotated corpora. We have trained classifiers based on the Conditional Random Fields algorithm (Lafferty *et al.*, 2001) for both disorder and chemical entity mentions. We will train more classifiers to recognize gene and protein mentions, with existing corpora annotated with those types of entities. The results of these classifiers will be evaluated by the crowd, who will be able to accept the entity and its boundaries, adjust the boundaries, or reject the entity if it does not correspond at all to what the classifier predicted. These corrections will be used to improve the performance of the first step, through reinforcement learning, with different weights assigned to the specialists according to their usage profile.

The normalization module will first attempt to map the string to a concept of the respective ontology. Since multiple matches may exist for the same string, this ambiguity will be solved with a semantic similarity measure. These mappings will be evaluated by the crowd, why the option of accepting the concept as correct, or choosing another one from the same ontology. As before, these corrections will be used to train a machine learning classifier, using the semantic similarity values as features.

For example, taking the sentence "The rhythm appears to be atrial fibrillation" as input, the NER classifier may recognize only the word "fibrillation" as a disorder mention. In this case, the boundary of the term may be extended to include "atrial". In SNOMED-CT, several concept are related to atrial fibrillation, for example, "Atrial fibrillation" (C0004238) and "Atrial fibrillation and flutter" (C0155709). If the second concept is chosen by the system instead of the first one, the user may indicate this mistake. Otherwise, the user will confirm that the mapping is correct.

Every document processed by our system is anonymized using standard procedures, which includes removing all references to personal details. The user only evaluates individual phrases containing annotations, to prevent the re-identification of documents. We will apply a sliding-window approach to harmonize the evaluations performed by the crowd, so that each phrase evaluated by a user should overlap with other phrases. With this strategy, we can align the sequence of phrases that was accepted by the majority of the crowd, and prevent errors committed due to the lack of context.

As an incentive to the participation of users, we intend to apply a mechanism of rewards based on a virtual currency. KnowledgeCoin (Couto, 2014) is a virtual currency that was originally proposed to reward and recognize data sharing and integration on the semantic web. This could also be applied to the proposed pipeline, by distributing KnowledgeCoins for each text validated by a user, improving the reputation of that user. Potential participants in this kind of project would be medicine students. The University of Lisbon accepts almost three hundred medicine students per year, which could provide a relatively large crowd for our pipeline. Retired physicians, nurses, physician assistants and researchers may also participate, in order to provide more specialized curation. This type of crowd has been used by CrowdMed to provide crowd-sourced diagnostics to complex medical cases, with high levels of accuracy.

5 CONCLUSION

We propose a novel pipeline for recognition and normalization of biomedical terms to ontology concepts, using crowd-sourcing. The complete and automatic annotation of biomedical texts such as EHR requires systems with high precision. The normalization task is particularly challenging due to the subjective nature of ontology mapping. By letting a large group of specialized participants correct the mistakes of a machine learning classifier, we expect an improvement of the performance of current biomedical text mining systems. The idea is not only to create a scalable knowledge base but help from a community of specialist curators that may be available to help in creating a golden standard for a new biomedical area, or improve current results, or just validate some results.

ACKNOWLEDGEMENTS

This work was supported by the Fundação para a Ciência e a Tecnologia (https://www.fct.mctes.pt/) through the PhD grant PD/BD/106083/2015, the Biosys PhD programme and LaSIGE Unit Strategic Project, ref. PEst-OE/EEI/UI0408/2014.

REFERENCES

Ambati, V. and Vogel, S. (2010). Can crowds build parallel corpora for machine translation systems? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 62–65. Association for Computational Linguistics.

Campbell, K. E., Oliver, D. E., and Shortliffe, E. H. (1998). The unified medical language system toward a collaborative approach for solving terminologic problems. *Journal of the American Medical Informatics Association*, **5**(1), 12–16.

Good, B. M. and Su, A. I. (2013). Crowdsourcing for bioinformatics. *Bioinformatics*, page btt333.

Jansen, D., Alcala, A., and Guzman, F. (2014). Amara: A sustainable, global solution for accessibility, powered by communities of volunteers. In *Universal Access in Human-Computer Interaction. Design for All and Accessibility Practice*, pages 401–411. Springer.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lamurias, A., Ferreira, J. D., and Couto, F. M. (2015). Improving chemical entity recognition through h-index based semantic similarity. *Journal of Cheminformatics*, **7**(Suppl 1), S13.

Mrozinski, J., Whittaker, E., and Furui, S. (2008). Collecting a why-question corpus for development and evaluation of an automatic qa-system. In *46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies*, pages 443–451.

Sarasua, C., Simperl, E., and Noy, N. F. (2012). Crowdmap: Crowdsourcing ontology alignment with microtasks. In *The Semantic Web–ISWC 2012*, pages 525–541. Springer.

Sukharev, J., Zhukov, L., and Popescul, A. (2014). Learning alternative name spellings on historical records.

Von Ahn, L. and Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, **51**(8), 58–67.