Can the Wisdom of the Crowd Be Used to Improve the Creation of Gold-standard for Text Mining applications?

Luis F. Campos, Andre Lamurias, Francisco M. Couto

LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

Abstract. Natural Language Processing (NLP) and text mining techniques require annotated datasets to develop and evaluate new approaches. These datasets are commonly developed by domain experts, who manually annotate a corpus of documents with the relevant information. This process is expensive and time-consuming, and for this reason, it has been suggested that crowdsourcing techniques could be used to develop a gold standard, with similar quality.
This paper presents a workflow for using the wisdom of the crowd to develop an annotated corpus. Our approach is focused on the annotation of named entities in biomedical text, using a crowd of non-experts. This workflow is adaptable to various types of text and domains. We tested this workflow on two occasions, one to compare the crowd annotations with a gold standard created by experts and another to annotate clinical reports with radiology terms.

Keywords: wisdom of the crowd; text mining; named-entity recognition

1 Introduction

Natural Language Processing (NLP) and text mining techniques require annotated datasets to develop and evaluate new approaches. These datasets consist of a corpus of documents relevant to a specific domain and its annotations, which can serve as a gold standard to train a machine learning classifier and evaluate its performance. These annotations are made by domain experts and require a defined set of annotation guidelines and time to read and annotate the texts. Recent crowdsourcing techniques have shown that non-experts can provide reliable annotations for text mining tasks. Snow et. al. [1] compared expert annotations with non-expert annotations and obtained a high agreement between the two on 5 different NLP tasks. Furthermore, they have shown that only a small number of non-expert annotators is necessary to achieve expert quality. Li et. al. [2] used crowdsourcing to annotate documents with relations between chemical compounds and diseases. Each chemical-disease pair was reviewed by 5 annotators. Their approach achieved an F-score of 0.505, while machine learning systems trained on a gold standard annotated by experts obtained an F-score of 0.570.

Our objective was to develop a flexible crowdsourcing workflow for annotation of biomedical concepts in text. To demonstrate the flexibility and feasibility of the workflow, we performed two experiments consisting of crowd annotation sessions on two different domains (phenotypes and radiology terms). With one of the sessions, we compared the quality of the crowd annotations with automatic annotations, with the other we applied the workflow to obtain a new annotated dataset. For our purposes, we define crowd as any number of people performing the same task with a common objective. The number of people necessary depends on the amount of text to be annotated in order to reach an acceptable level of quality. Both sessions were organized in the context of a classroom, i.e., the annotation session occurred in the same space during a fixed period of time. This way, we could explain clearly the motivation and objectives of the experiment and solve any technical issues occurring during the annotation process. This paper describes the crowdsourcing workflow we developed for annotation of biomedical documents, as well as the specifics of each annotation session. Furthermore, we present the results obtained with each experiment, a discussion about the advantages of this approach and further improvements.

2 Methods

In this section, we will first explain in general terms the idea of using the wisdom of the crowd to develop a gold standard and then focus on the two experiences we had implementing it. A crowd annotation session consists of a period of time where human annotators are asked to annotate documents according to some guidelines. Our goal with a crowd annotation session is to get annotations of terms of a certain domain on a set of documents identified by the session participants. To accomplish this goal, it is necessary to carefully prepare each annotation session.

The first step should be to define the domain and a respective set of documents relevant to that domain. The source of these documents will depend on the nature of the experiment: if the objective is to evaluate crowd annotations on a specific domain, a corpus with gold standard annotations should be chosen, so that the crowd annotations can be evaluated against the gold standard. However, if the objective is to develop a new gold standard, or to improve automatic annotations, it is necessary to select an appropriate set of documents to annotate. There are several source of biomedical text that can be used to this end, such as PubMed and ClinicalTrials.gov. The number of documents should be chosen according to duration of the sessions. Since the participants are not domain experts, they will have to spend time reading the guidelines and verifying the definition of various concepts.

After defining the topic and corpus, it is necessary to define the type of information to be annotated. In our case, we focused on defining named entity annotations, which consists of identifying the words that refer to entities of a specific domain. It should be clear to the participants what should be annotated in the text. Hence, an existing vocabulary or ontology can be used, so that the

annotators have plenty examples of the entities. In our experiments, we used two ontologies that provide a web interface that can be used to search for terms, and most terms contain a definition. This way, the annotators can confirm if an entity should be annotated or not based on its definition. Other criteria should also be defined, similar to the annotation guidelines used for other corpus annotation projects. Krallinger et. al. [3] provides a detailed description of the guidelines used to annotated chemical compounds by experts. In our case, the guidelines should be simple to understand by non-experts. As such, we compiled a list bullet points containing the main principles that the annotators could consult during the session.

It is necessary to choose how the annotators can access the text and submit the annotations. There are various frameworks that can be used to this purpose. Commercial frameworks, such as CrowdFlower and Amazon Mechanical Turk, provide more options and features and can recruit annotators from a large pool of users, which are paid for each annotation or document. However, there are other examples of volunteer-based crowdsourcing projects, such as Folding@Home [4] and Mark2Cure [5]. We used WebAnno [6] for the two experiments because it was freely available and adaptable to our needs; i.e. it was possible to upload documents and organize them in projects, and the annotators could access the documents and directly annotate each document. Then, we could export the annotations of each user and analyze them with our scripts. During the sessions, we demonstrate to the users the annotation guidelines and how they should use the annotation platform and the guidelines for doing it. To bootstrap the annotation process we automatically pre-annotate the corpus with terms of the domain, based on an existing vocabulary. This way, the annotators can accept and reject the automatic annotations, or add new ones.

2.1 Evaluation metrics

After each annotation session, we show and discuss the results with the participants. We chose to analyze the results right after the end of the session and share with the participants, to emphasize the importance of their work. For the HPO experiment, for which we had a gold standard, we calculated precision, recall, and F1-score. For both experiments, we compared the annotations of the participants based on an agreement score. The agreement score measures how agreeable is each participant, comparing the annotations of the majority of the participants. We wanted to obtain a high score to the participants that voted most times like the majority, and a low score to participants that voted differently more often. By ranking the participants by this agreement score, we could filter out noisy annotations. We defined the agreement score of a participant as the sum of the fraction of users who agreed with the participant on each annotation. Since we provided automatically generated annotations for each document, each participant could either accept or reject each annotation:

$$ballot_{a,p} = \begin{cases} 1 & \text{if participant } p \text{ validated annotation } a \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where p is a participant and a an automatic annotation. We can compare the ballot of two users on the same annotation:

$$v(a, p1, p2) = \begin{cases} 1 & \text{if } ballot_{a,p1} = ballot_{a,p2} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Hence, we calculated the agreement score with the following formula:

$$agreement\_score(p) = \sum_{a \in A} \frac{1}{\#P} \sum_{i \in P} v(a, p, i) \tag{3}$$

where A is the set of automatic annotations and P is the set of all participants. For example, considering an annotation accepted by 70% of the participants, means that those who accepted receive a score of 0.7 for that annotation, while the ones who rejected it would receive score of 0.3. Hence, this score is related to how much the annotations made by the participant have in common with the crowd.

The agreement score only takes into account how the crowd performed on pre-existing annotations, even though they could also add new annotations. We decided to not include new annotations in the agreement score since it is a different task from simply reviewing automatic annotations, comparing to searching actively for missing annotations. We assigned a novelty score to each participant, which measures the number of novel annotations added by each user:

$$novelty\_score(p, t) = \sum_{a \in A} ballot(p, a) \text{ if } \frac{1}{\#P} \sum_{i \in P} ballot(i, a) \geq t \tag{4}$$

where A is the set of annotations added by the participants. We can change the threshold t whether we want to reward novel annotations regardless if they were also added by other participants. The agreement and novelty scores can be used to rank the participants of the annotation session and study the differences within a crowd of annotators.

2.2 HPO terms annotation

In this annotation session, the participants had to annotate a specific part of a corpus [7] with HPO (Human Phenotype Ontology) [8] terms. The participants of this session were a group of 13 students enrolled in a systems biology PhD program. The gold standard annotations of this corpus allowed us to compare the crowd annotations with the gold standard. We used MER [9] to obtain baseline annotations which participants could accept or reject, or also add new annotations. Six documents from this corpus were annotated during approximately 40 minutes. Material related to this session can be found at https://github.com/lasigeBioTM/HPO_Crowd_Annotation_Experiment.

2.3 Radiology terms annotation

In the context of a biomedical text mining workshop, we organized hands-on activity consisting of the annotation of RadLex terms [10] on Radiology reports by the workshop participants. We selected three Radiology reports from Lurie Children's TF Library http://mirc.luriechildrens.org/query. We performed an informal preliminary test to understand how long it would take to annotate each report and other issues that could occur. Since we had limited time to perform this activity, we decided to select 3 documents for this session, expecting an average of 10 minutes per document. We also used the NCBO annotator [11] to present automatic annotations to the participants. The rationale for this activity was that the aggregate annotations of the participants would be used as a wisdom of the crowd gold standard, which could be used to train and evaluate machine learning approaches. Material related to this session can be found at https://github.com/lasigeBioTM/biomedical_workshop_bod.

3 Results and Discussion

3.1 HPO annotation session

During the HPO annotation session, we obtained annotations of six documents from 13 participants. The crowd annotations were created as described in the previous section and compared with the gold standard annotations developed by domain experts. Here the rationale was to check if the wisdom of the crowd could replace the use of (expensive) experts in the development of gold standard annotations. We experimented with a number of validation thresholds to study which one works best. This validation threshold corresponds to the fraction of participants that have to agree on an annotation to accept it.

Using different validation thresholds, we observe that the precision values tends to increase with the threshold, while the opposite happens to the recall (see Table 3.1 and Figure 3.1). The maximum F1-score is obtained with a threshold of 0.6 (0.659) while the maximum precision is obtained with a threshold of 0.7 (0.815) and maximum recall with a threshold of 0.1 (0.868). The plot in Figure 3.1 is similar to the analogous ones presented in [12, 5], which describe similar experiments, but on a larger scale (more participants and using more documents) and in a different domain (disease mentions). These differences might explain why in our case we obtained worse F1-Scores.

The F1-score obtained by our baseline (MER annotations) on the same 6 documents was 0.618 (Table 3.1. Using the crowd with a threshold of 0.6, we obtained a higher F1-score (0.659). These results show that the crowdsourcing approach outperformed a rule-based system, especially in terms of precision. We have previously developed IHP [1], a system based on machine learning and post-processing rules to recognize human phenotypes. This system was evaluated with cross-validation on the full HPO corpus. We considered only the same 6

[1] https://github.com/lasigeBioTM/IHP

Threshold	Precision	Recall	F1-Score
0.1	0.315	**0.868**	0.462
0.2	0.429	0.849	0.462
0.3	0.500	0.774	0.607
0.4	0.547	0.660	0.598
0.5	0.660	0.623	0.641
0.6	0.789	0.566	**0.659**
0.7	**0.815**	0.415	0.550
0.8	0.800	0.302	0.438
0.9	0.786	0.201	0.328

Table 1. Metrics evaluating the quality of the crowd annotations against gold standard annotation, using different aggregation thresholds, e.g., the 0.5 row means the only annotations accepted by at least 50% of the participants were included in the crowd annotations.

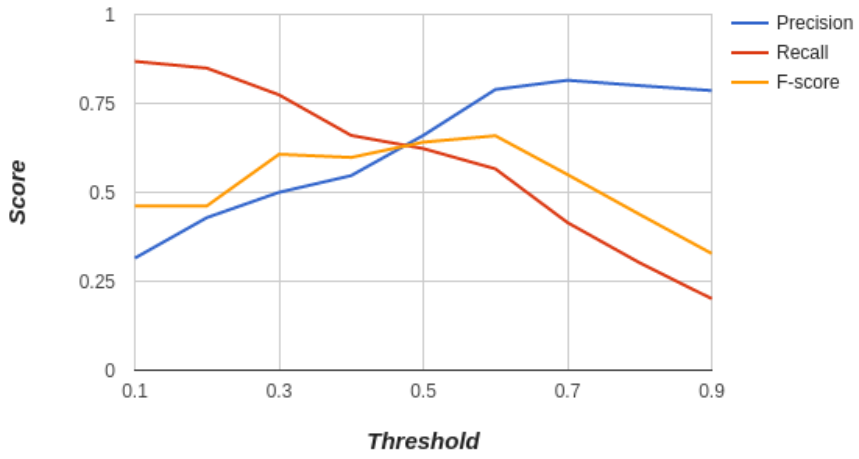

Fig. 1. Graphical representation of the scores obtained with various validation thresholds.

documents and calculated the same scores, to compare with the scores obtained by the crowd. While the crowd obtained higher precision, IHP obtained a higher recall, resulting in a higher F1-score. However, it should be taken into account that IHP was highly tuned for this corpus, using the gold standard annotations and manually tuned post-processing rules, while the crowd annotations were obtained in approximately 40 minutes. Furthermore, a larger number of documents may be necessary to perform a stronger comparison.

	Precision	Recall	F1-Score
MER	0.596	0.642	0.618
Crowd	**0.789**	0.566	0.659
IHP	0.577	**0.849**	**0.687**

Table 2. Comparison between a baseline approach (MER), crowd annotations, using the highest F-score threshold (0.6), and machine learning and post-processing rules (IHP).

3.2 RadLex annotation session

The RadLex annotation session resulted in 3 radiology reports annotated by 25 participants. We created a crowd annotation file for each of the radiology reports being used by including in the crowd annotation file just the annotations that were done by at least 80% of session participants. Since we did not have expert gold standard annotations for this corpus, we analyzed manually the effect of the crowd on the baseline annotations. We observed that the crowd was able to improve the specificity of the annotations in some cases. While the baseline annotations contained "left retinoblastoma" as separate entities, most of the participants merged the two entities. In other cases, the crowd added more words to an entity, changing "tumor" to "pineal tumor", for example.

Another case that demonstrates the potential of the crowd was the identification of entities with typos. In one of the reports, the word "adolsecent" is used, and it is clear that the author meant "adolescent", a radiology term according to RadLex. Automatic approaches to detect typos are limited in terms of performance, and vocabularies to do not include this type of variation. The fact that the crowd was able to detect this issue is another advantage of this approach.

3.3 Comparison between annotation sessions

We calculated the agreement score for each participant of the two annotation sessions and normalized the value by dividing by the total number of annotations. The maximum agreement score obtained during the HPO session was 0.730 while for the RadLex session it was 0.832. While the radiology reports may be more difficult to understand for non-experts, the HPO vocabulary seemed to be more ambiguous. During the HPO session the participants had more doubts about what was relevant, resulting in more disagreements in the annotations. Comparing the ranking of participants, we noticed that some participants obtained both high agreement and novelty scores. In the HPO session, the top 5 of both rankings had 3 participants in common, while in the RadLex session had, there were 2 participants in common. We also identified one participant in the RadLex session who was in the top 5 of the agreement score ranking and in the bottom 5 of the novelty score. This case can have various explanations: the participant could have had technical difficulties in adding new annotations, could

have been more focused on validating the automatic annotations, or simply did not think that many annotations were missing.

4 Conclusion

In this paper, we present our workflow for crowdsourcing annotations using non-experts. We demonstrated this workflow on two sessions: one where a gold standard was available and it was possible to evaluate the crowd annotations, and another where no gold standard was available. With the first dataset we obtained an F1-score superior to a dictionary matching approach and using different validation thresholds, we can obtain higher precision or recall values. With the second dataset, we demonstrated how the workflow can be applied to different domains and types of text. In this case, the documents consisted of clinical reports, and the crowd was able to improve the quality of the baseline annotations.

The experiments presented in this paper represent a proof-of-concept for future, more extensive, crowd annotation projects. While in both experiments every participant annotated every document, some authors suggest that only a few annotators are necessary to annotate a document [2]. This way, using the same number of participants and in the same time duration, we could obtain a larger number of annotated documents. More complex text mining tasks can also be performed using the wisdom of the crowd [13]. Named entity normalization and relation extraction are two tasks that can be attempted in the future using our workflow [14, 15]. This workflow can also be integrated with active learning algorithms to select the best annotators for each document [16].

Acknowledgments. This work was supported by the Fundação para a Ciência e a Tecnologia (https://www.fct.mctes.pt/) through the PhD Grant ref. PD/BD/106083/2015 and UID/CEC/00408/2013 (LaSIGE).

References

1. Snow, R., Connor, B.O., Jurafsky, D., Ng, A.Y., Labs, D., St, C.: Cheap and fast - but is it good? Evaluation non-expert annotiations for natural language tasks. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (October) (2008) 254–263
2. Li, T.S., Bravo, A., Furlong, L.I., Good, B.M., Su, A.I.: A crowdsourcing workflow for extracting chemical-induced disease relations from free text. Database (Oxford) (2016) 1–11
3. Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D.M., Sayle, R.A., Batista-Navarro, R.T., Rak, R., Huber, T., Rocktschel, T., Matos, S., Campos, D., Tang, B., Xu, H., Munkhdalai, T., Ryu, K.H., Ramanan, S.V., Nathan, S., Zitnik, S., Bajec, M., Weber, L., Irmer, M., Akhondi, S.A., Kors, J.A., Xu, S., An, X., Sikdar, U.K., Ekbal, A., Yoshioka, M., Dieb, T.M., Choi, M., Verspoor, K., Khabsa, M., Giles, C.L., Liu, H., Ravikumar, K.E., Lamurias, A., Couto, F.M., Dai, H.J., Tsai, R.T.H., Ata, C., Can, T., Usi,

A., Alves, R., Segura-Bedmar, I., Martinez, P., Oyarzabal, J., Valencia, A.: The CHEMDNER corpus of chemicals and drugs and its annotation principles. Journal of Cheminformatics **7**(Suppl 1) (2015) 1–17
4. Beberg, A.L., Ensign, D.L., Jayachandran, G., Khaliq, S., Pande, V.S.: Folding@ home: Lessons from eight years of volunteer distributed computing. In: Parallel & Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on, IEEE (2009) 1–8
5. Tsueng, G., Nanis, M., Fouquier, J., Good, B., Su, A.: Citizen Science for Mining the Biomedical Literature. bioRxiv (2016) 038083
6. Eckart de Castilho, R., Mujdricza-Maydt, E., Yimam, S.M., Hartmann, S., Gurevych, I., Frank, A., Biemann, C.: A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In: Proceedings of the workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH) at COLING 2016. (2016) 76–84
7. Groza, T., Köhler, S., Doelken, S., Collier, N., Oellrich, A., Smedley, D., Couto, F.M., Baynam, G., Zankl, A., Robinson, P.N.: Automatic concept recognition using the human phenotype ontology reference and test suite corpora. Database (Oxford) (2 2015) 1–13
8. Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., Brudno, M., Buske, O.J., Chinnery, P.F., Cipriani, V., Connell, L.E., Dawkins, H.J.S., DeMare, L.E., Devereau, A.D., de Vries, B.B.A., Firth, H.V., Freson, K., Greene, D., Hamosh, A., Helbig, I., Hum, C., Jähn, J.A., James, R., Krause, R., Laulederkind, S.J.F., Lochmüller, H., Lyon, G.J., Ogishima, S., Olry, A., Ouwehand, W.H., Pontikos, N., Rath, A., Schaefer, F., Scott, R.H., Segal, M., Sergouniotis, P.I., Sever, R., Smith, C.L., Straub, V., Thompson, R., Turner, C., Turro, E., Veltman, M.W.M., Vulliamy, T., Yu, J., von Ziegenweidt, J., Zankl, A., Züchner, S., Zemojtel, T., Jacobsen, J.O.B., Groza, T., Smedley, D., Mungall, C.J., Haendel, M., Robinson, P.N.: The Human Phenotype Ontology in 2017. Nucleic Acids Research **45**(Database) (1 2017)
9. Couto, F.M., Campos, L.F., Lamurias, A.: MER: a Minimal Named-Entity Recognition Tagger and Annotation Server. Proceedings of the BioCreative V.5 Challenge Evaluation Workshop (2017) 130–137
10. Langlotz, C.P.: RadLex: a new method for indexing online educational materials. RadioGraphics **26**(6) (11 2006) 1595–1597
11. Jonquet, C., Shah, N., Musen, M.: The open biomedical annotator. In: AMIA summit on translational bioinformatics. (2009) 56–60
12. Good, B.M., Nanis, M., Wu, C., Su, A.I.: Microtask Crowdsourcing for Disease Mention Annotation in Pubmed Abstracts. Biocomputing 2015 (2014) 282–293
13. Lamurias, A., Pedro, V., Clarke, L., Couto, F.: Annotating biomedical terms in electronic health records using crowd-sourcing. In: International Conference on Biomedical Ontologies (ICBO), Early Career. (2015)
14. Lamurias, A., Ferreira, J.D., Couto, F.M.: Improving chemical entity recognition through h-index based semantic similarity. Journal of cheminformatics **7**(S1) (2015) S13
15. Lamurias, A., Clarke, L.A., Couto, F.M.: Extracting microRNA-gene relations from biomedical literature using distant supervision. PloS one **12**(3) (2017) e0171929
16. Fang, M., Yin, J., Tao, D.: Active learning for crowdsourcing using knowledge transfer. Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence **3** (2014) 1809–1815