Community annotation experiment for ground truth generation for the i2b2 medication challenge

Özlem Uzuner,[1] Imre Solti,[2] Fei Xia,[3] Eithon Cadag[2]

► An additional material is published online only. To view this file please visit the journal online (http://jamia.bmj.com).

[1]Department of Information Studies, University at Albany, State University of New York, Albany, New York, USA
[2]Division of Biomedical and Health Informatics, University of Washington, Seattle, Washington, USA
[3]Department of Linguistics, University of Washington, Seattle, Washington, USA

Correspondence to
Özlem Uzuner, College of Computing and Information, University at Albany, SUNY, Draper 114A, 135 Western Ave, Albany, NY 12222, USA; ouzuner@albany.edu

Received 20 February 2010
Accepted 25 June 2010

ABSTRACT

Objective Within the context of the Third i2b2 Workshop on Natural Language Processing Challenges for Clinical Records, the authors (also referred to as 'the i2b2 medication challenge team' or 'the i2b2 team' for short) organized a community annotation experiment.

Design For this experiment, the authors released annotation guidelines and a small set of annotated discharge summaries. They asked the participants of the Third i2b2 Workshop to annotate 10 discharge summaries per person; each discharge summary was annotated by two annotators from two different teams, and a third annotator from a third team resolved disagreements.

Measurements In order to evaluate the reliability of the annotations thus produced, the authors measured community inter-annotator agreement and compared it with the inter-annotator agreement of expert annotators when both the community and the expert annotators generated ground truth based on pooled system outputs. For this purpose, the pool consisted of the three most densely populated automatic annotations of each record. The authors also compared the community inter-annotator agreement with expert inter-annotator agreement when the experts annotated raw records without using the pool. Finally, they measured the quality of the community ground truth by comparing it with the expert ground truth.

Results and conclusions The authors found that the community annotators achieved comparable inter-annotator agreement to expert annotators, regardless of whether the experts annotated from the pool. Furthermore, the ground truth generated by the community obtained F-measures above 0.90 against the ground truth of the experts, indicating the value of the community as a source of high-quality ground truth even on intricate and domain-specific annotation tasks.

INTRODUCTION

Ground truth forms the basis of all natural language processing (NLP) research. Traditionally, ground truth is generated by a team consisting of guideline designers, annotators, and technical support staff. The annotators in this team are either domain experts or are trained on the annotation task for a long period of time[1–5]; therefore, we refer to their annotations as 'expert annotations'.

Expert annotations require recruitment and training of usually a small number of experts and the execution of the actual annotation; therefore, their generation takes significant funding and time. We hypothesize that annotations generated by the community can provide a viable alternative to expert annotations. In order to test this hypothesis, the authors (also referred to as 'the i2b2 medication challenge team' or 'the i2b2 team') organized a community annotation experiment and studied the quality of community annotations.

This experiment addressed the task set by the Third i2b2 Workshop on Natural Language Processing Challenges for Clinical Records. The goal was the extraction of information on medications of patients from discharge summaries. The workshop was therefore referred to as the medication challenge. The information to be annotated for the medication challenge was classified into seven 'fields': medication names, their doses, modes (routes) of administration, frequencies and durations of administration, the reasons for administering each medication, and whether the medication was mentioned in a list or in the narrative running text of the discharge summary. The medication challenge asked that the set of field values that relate to a specific mention of a medication be linked together to create an 'entry' if the field values were specified within two lines of the medication mention. Field values mentioned outside of the two-line window were considered out of scope, and their values were set to 'nm' for 'not mentioned'.

Seventy-nine individuals from 20 teams contributed to the community annotation experiment. In this article, we describe this experiment and discuss its results. Details of the systems developed for the medication challenge and their evaluation can be found in Uzuner *et al*.[6]

BACKGROUND AND RELATED WORK

The increasing availability of annotated corpora has fueled the advancement of NLP in both medical and open-domain language processing. Syntactic (eg, the English Penn Treebank[7] and the PARC 700 Dependency Bank[8]) and semantic/discourse annotations of corpora (eg, PropBank[9 10] and the Penn Discourse Treebank[11]) have supported the development of NLP systems and enabled their head-to-head comparison. Shared tasks organized on these corpora have fostered creativity, collaboration, and community spirit across the field.[12 13]

While most of the currently available corpora are annotated by experts, the last few years have seen an increase in the efforts to acquire annotations by taking advantage of online labor markets such as Amazon's Mechanical Turk (AMT).[14 15] We believe that annotations can be gathered through AMT when the task is relatively simple and does not require any domain knowledge. However, our annotation task is quite intricate: it involves

extracting medications and medication-related information from discharge summaries. This requires that the annotators be well-trained in the task. Given limited funding and time for the medication challenge (which made expert annotation a remote possibility) and the complexity of the task (which made the use of AMT very risky), the most appropriate annotators come from the community. The community involved in this experiment consists of the medication challenge participants because of their vested interest in the challenge.

DATA

The data for the medication challenge consisted of 1243 deidentified discharge summaries obtained from Partners Healthcare: 547 of these discharge summaries were released to the challenge participants as training data (17 of which were annotated by the i2b2 medication challenge team), and the remaining 696 summaries were used as test data. These summaries were released to the challenge participants with the understanding that participation in the challenge constituted commitment on the part of the challenge teams to contribute to the community annotation experiment. In total, 251 discharge summaries from the test data were annotated by the challenge teams (after all teams submitted their system outputs to i2b2). All relevant institutional review boards approved the challenge and the use of the discharge summaries for research.

METHODS

The community annotation experiment unfolded in three phases: in the first phase, we generated the annotation guidelines; in the second phase, the community annotated discharge summaries for the challenge; and in the third phase, we evaluated the quality of the community annotations and agreement.

GENERATION OF ANNOTATION GUIDELINES

The medication challenge guidelines assumed no medical, linguistic, or computer science background on the part of the annotators. They also did not assume the annotators to be native English speakers. These guidelines were created through an iterative process during which a group of students at the University of Washington annotated a small set of discharge summaries based on the guidelines, measured inter-annotator agreement, and posed questions that helped us revise the guidelines. After several iterations, the annotation guidelines and 17 annotated discharge summaries were released to the challenge teams.

During the next 3 weeks, the challenge teams studied the annotation guidelines and the annotated discharge summaries, asked clarifying questions, and helped address any residual inconsistencies found in the annotations of the 17 discharge summaries. At the end of this period, the i2b2 team froze the annotation guidelines and continued to answer clarifying questions which required interpretation of the guidelines in light of new examples.

ANNOTATION GUIDELINES

The medication challenge annotation guidelines defined the following 'fields':

1. Medications (m): included names, brand names, generics, and collective names of prescription substances, over-the-counter medications, and other biological substances required or suggested by doctors, for which the patient is the experiencer—for example, Lasix, aspirin, total prenatal nutrition.
2. Dosages (do): referred to the amount of a single medication used in each administration—for example, one tab, 4 units, 30 mg.
3. Modes (mo): referred to the route for administering the medication—for example, oral, intravenous, topical.
4. Frequencies (f): referred to how often each dose of the medication should be taken—for example, daily, ×1, once a month, 3 times a day.
5. Durations (du): referred to how long the medication is to be administered—for example, for a month, during spring break, until the symptom disappears.
6. Reasons (r): referred to the medical reason for which the medication is stated to be given—for example, fever, diabetes.
7. List/narrative (ln): marked whether the medication information appears in a list structure or in narrative running text in the discharge summary.

Any information related to medications that were not experienced by the patients was excluded from the medication challenge.

ANNOTATION PROCESS

For the community annotation experiment, 251 discharge summaries were allocated to the challenge teams. A subset of these summaries was also annotated by the i2b2 team in order to provide expert annotations which can be compared with the community annotations.

COMMUNITY ANNOTATIONS

The community annotation took place after system outputs were turned into i2b2 and was conducted in two phases: initial annotation and adjudication. Each challenge team was allocated 10 discharge summaries per person, with some relief provided for any training discharge summary annotations (the team may have optionally developed during training) that the team would share with the i2b2 for inclusion with the challenge data for future research. Each team's annotation allocation was split between initial annotation and adjudication, if possible.

Each discharge summary was assigned to two independent challenge teams for initial annotation. Along with their allocated discharge summaries, each challenge team was provided with the pooled system outputs for those summaries. The pooled system outputs were obtained by polling the system outputs for completeness. The three most densely populated system outputs submitted by three different teams were used as the pool for each discharge summary. The initial annotation took 2 weeks.

After initial annotation, the i2b2 team automatically merged the initial annotations so that any annotations that the two teams agreed on were added to the 'penultimate ground truth'. The disagreements between the two teams were passed on to a third challenge team for adjudication. Adjudication took 2 weeks.

After adjudication, conflicting annotations were checked for validity by the i2b2 team and, if approved, they were added to the penultimate ground truth. The penultimate ground truth was released to the community for scrutiny and for suggestions for improvement. During a 3-week period after the release of the penultimate ground truth, community corrections were vetted and included in the ground truth if agreed upon by the i2b2 team. The community-corrected ground truth was considered final. We refer to this ground truth as the 'final community ground truth'. All system evaluations in the medication challenge were run against the final community ground truth.

Table 1 Two datasets used for comparing community and expert annotations; the number of entries and fields for these datasets

		Final community ground truth				Expert ground truth			
	# of discharge summaries		# of entries	Fields	# of instances per field		# of entries	Fields	# of instances per field
Set A	24	From pooled system outputs	867	m	867	From pooled system outputs	913	m	913
				do	450			do	470
				mo	333			mo	345
				f	434			f	448
				du	59			du	62
				r	125			r	141
Set B	24	From pooled system outputs	747	m	747	From raw discharge summaries	766	m	766
				do	342			do	351
				mo	246			mo	248
				f	287			f	287
				du	43			du	43
				r	182			r	188

do, dosages; du, durations; f, frequencies; m, medications; mo, modes; r, reasons.

EXPERT ANNOTATIONS

In order to provide annotations that can represent traditionally generated ground truth, the i2b2 medication challenge team annotated two non-overlapping subsets of the 251 test discharge summaries (table 1). The first set, set A, consisted of 24 discharge summaries that the i2b2 team annotated on the basis of pooled system outputs. Two i2b2 team members provided initial annotations on these summaries, and a third member adjudicated. Their pooled system outputs were identical with the pooled outputs used by the challenge participants for annotating the same summaries.

In order to test the effect of pooled system outputs on inter-annotator agreement, the i2b2 team annotated another set, set B, of 24 discharge summaries from the 251 test summaries from scratch. Two of the i2b2 team members provided initial annotations on these summaries, and a third adjudicated disagreements.

METRICS OF AGREEMENT

We measured the agreement[16 17] between annotators using horizontal and vertical F-measures. Horizontal F-measures are computed over 'entries', whereas vertical F-measures are computed over individual 'fields'. Both sets of F-measures are computed at both phrase and token level. For this purpose, a phrase is the exact text corresponding to the value of an individual field, and a token is an individual word in the text of the field. Equations show phrase- and token-level precision, recall, and F-measures:

$$\text{Phrase-level precision (PP)} = \frac{\#\text{ Correctly returned phrases by system}}{\#\text{ phrases returned by the system}} \quad \text{Equation 1}$$

$$\text{Phrase-level recall (PR)} = \frac{\#\text{ Correctly returned phrases by system}}{\#\text{ phrases in gold standard}} \quad \text{Equation 2}$$

$$\text{Phrase-level F-measure (PF)} = \frac{\left(\beta^2+1\right)\times \text{PP}\times \text{PR}}{\left(\beta^2\times \text{PP}\right)+\text{PR}} \text{ where } \beta=1 \quad \text{Equation 3}$$

where β marks the relative weights of precision and recall.

$$\text{Token-level Precision (TP)} = \frac{\#\text{ Correctly returned tokens from each phrase in system output}}{\#\text{ tokens in system output}} \quad \text{Equation 4}$$

$$\text{Token-level Recall (TR)} = \frac{\#\text{ Correctly returned tokens from each phrase in system output}}{\#\text{ tokens in ground truth}} \quad \text{Equation 5}$$

$$\text{Token-level F-measure (TF)} = \frac{\left(\beta^2+1\right)\times \text{TP}\times \text{TR}}{\left(\beta^2\times \text{TP}\right)+\text{TR}} \text{ where } \beta=1 \quad \text{Equation 6}$$

Micro-averaged performance metrics aggregate the entries from all the discharge summaries in the test data and compute horizontal and vertical metrics over all the entries; macro-averaged metrics are computed per individual discharge summary and are then averaged. In the next section, we use macro-averaged F-measures to compute agreement between annotators, and micro-averaged F-measures to compute the difference between two sets of ground truth. We measure the significance of the difference in F-measures using approximate randomization (see online supplements).[18 19]

RESULTS AND DISCUSSION

We measure agreement between pairs of annotators on creating the ground truth. Agreement is a measure of reliability.[17] We report agreement results for each of the ground truth sets separately. In order to measure quality of annotations, we compare the final community ground truth and the expert ground truth.

AGREEMENT AND RELIABILITY

Table 2 shows the macro-averaged horizontal F-measure of the initial community annotators when they annotated from pooled system outputs, of the expert annotators when they annotated from pooled system outputs, and of the expert annotators when they annotated from scratch. The differences in the macro-averaged F-measures of the various pairs of annotators were not statistically significant. Agreement, as measured by macro-

Table 2 Macro-averaged horizontal F-measures for pairs of annotators

Annotator 1 vs annotator 2		Macro-averaged horizontal F-measure
Community annotations from pooled system outputs on 251 test discharge summaries	Phrase level	0.824
	Token level	0.829
Expert annotations on set A	Phrase level	0.863
	Token level	0.880
Community annotations on set A	Phrase level	0.866
	Token level	0.870
Expert annotations on set B	Phrase level	0.846
	Token level	0.843
Community annotations on set B	Phrase level	0.812
	Token level	0.816

averaged horizontal F-measure, between all pairs of annotators was above 0.82.

Table 3 shows the macro-averaged vertical F-measures of the annotations of the experts and of the community annotators. The differences in the macro-averaged vertical F-measures were not significant for any of the fields. In other words, agreement between community annotators was comparable to the agreement between experts. In addition, compared with medication names, dosages, modes, and frequencies, the agreement on reasons and durations was much lower for both the expert and the community annotators, indicating that reasons and durations were difficult to identify precisely. Analysis of the major areas of disagreement between the annotators reveals that, in contrast to medication names which showed high inter-annotator agreement on complete phrases, reasons and durations contained many cases of partial agreement where the annotators disagreed on the boundaries of the phrases. We hypothesize that these disagreements were accentuated by the greater variability in the text used for reasons and durations, as well as the greater number of tokens and the greater variability in the number of tokens used in these fields.

Quality

To measure the difference between the final community ground truth and the expert generated ground truth, we computed micro-averaged F-measures.[6] Table 4 shows that the micro-averaged F-measures of the final community ground truth against the expert generated ground truth were above 0.90, indicating that the two sets of annotations are of similar quality.

Among the six fields extracted for the medication challenge, the F-measures for medications, dosages, modes, and frequencies were well above 0.90, whereas the F-measures for durations and reasons were much lower. This observation agrees with the lower agreement among all annotators on reasons and durations, indicating that these fields were difficult for all expert and community annotators.

Table 3 Macro-averaged vertical F-measures on individual fields

		Set A		Set B	
		Macro-averaged F-measures of experts	Macro-averaged F-measures of community	Macro-averaged F-measures of experts	Macro-averaged F-measures of community
Phrase level	m	0.868	0.878	0.875	0.847
	do	0.852	0.858	0.839	0.823
	mo	0.854	0.844	0.810	0.840
	f	0.859	0.871	0.863	0.823
	du	0.417	0.391	0.190	0.275
	r	0.478	0.366	0.329	0.322
Token level	m	0.905	0.906	0.906	0.863
	do	0.877	0.868	0.827	0.827
	mo	0.854	0.835	0.797	0.847
	f	0.872	0.884	0.847	0.815
	du	0.507	0.464	0.201	0.373
	r	0.522	0.475	0.384	0.437

No significant differences in agreement of experts from agreement of community annotators at p=0.05 and n=10 000.
do, dosages; du, durations; f, frequencies; m, medications; mo, modes; r, reasons.

Table 4 Micro-averaged F-measures of the final community ground truth against expert ground truth

	Set A		Set B	
	Phrase level	Token level	Phrase level	Token level
Overall (horizontal)	0.924	0.927	0.903	0.904
Medications (vertical)	0.925	0.943	0.909	0.939
Dosages (vertical)	0.956	0.965	0.938	0.946
Modes (vertical)	0.965	0.958	0.945	0.926
Frequencies (vertical)	0.957	0.965	0.957	0.959
Durations (vertical)	0.649	0.709	0.613	0.645
Reasons (vertical)	0.717	0.753	0.705	0.714

CONCLUSION

The medication challenge showed that, even on a relatively complex annotation task, community annotators can achieve inter-annotator agreement that is comparable to inter-annotator agreement of the experts. What is more, the ground truth obtained from community annotators agrees with the ground truth generated by expert annotators with horizontal F-measures above 0.90. These results justify the involvement of the community in ground truth generation and open up doors to annotation options that overcome the ground truth development bottleneck that is often encountered in NLP research.

Acknowledgments We thank the students in the Computational Linguistics Professional Master (CLMA) program at the University of Washington for their contributions to the development of the annotation guidelines, challenge teams for their contributions to the challenge, University of Washington and Harvard Medical School for their technical support, and AMIA for co-sponsoring the workshop.

Funding This work was supported in part by the NIH Roadmap for Medical Research, Grant U54LM008748, in addition to grants 2 T15 LM007442-06, 5 U54 LM008748, and 1K99LM010227-0110, and grants T15 LM07442, HHSN 272200700057 C, and N00244-09-1-0081. Other funders: NIH; NSF.

Ethics approval This study was conducted with the approval of the Partners Healthcare, SUNY, MIT, University of Washington.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Friedman C,** Hripcsak G, Shablinsky I. An evaluation of natural language processing methodologies. *Proceedings of AMIA* 1998:855—9.
2. **Uzuner Ö,** Luo Y, Szolovits P. Evaluating the State-of-the-Art in Automatic De-identification. *J Am Med Inform Assoc* 2007;**14**:550—63.
3. **Uzuner Ö,** Goldstein I, Luo Y, *et al*. Identifying Patient Smoking Status from Medical Discharge Summaries. *J Am Med Inform Assoc* 2008;**15**:14—24.
4. **Uzuner Ö.** Recognizing obesity and co-morbidities in sparse data. *J Am Med Inform Assoc* 2009;**16**:561—70.
5. **Pestian JP,** Brew C, Matykiewicz P, *et al*. A Shared Task Involving Multi-label Classification of Clinical Free Text. *Proceedings of BioNLP Workshop at the Annual Meeting of Association for Computational Linguistics* 2007:97—104.
6. **Uzuner Ö,** Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;**17**:514—18.
7. **Marcus MP,** Santorini B, Marcinkiewicz MA. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics* 1993;**19**:313—30.
8. **King TH,** Crouch R, Riezler S, *et al*. The PARC700 Dependency Bank. *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-2003)* 2003:1—8.
9. **Palmer M,** Gildea D, Kingsbury P. The proposition bank: an annotated corpus of semantic roles. *Computational Linguistics* 2005;**31**:71—106.

10. **Babko-Malaya O,** Bies A, Taylor A, *et al*. Issues in synchronizing the english treebank and PropBank. *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006 at the Annual Meeting of the Association for Computational Linguistics* 2006:70—7.
11. **Miltsakaki E,** Prasad R, Joshi AK, *et al*. The Penn Discourse TreeBank. *Proceedings of the Language Resources and Evaluation Conference (LREC-2004)* 2004.
12. **Tanabe L,** Xie N, Thom LH, *et al*. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics* 2005;**6**(Suppl 1):S3.
13. **Hersh W,** Cohen AM, Roberts P, *et al*. 2006 genomics track overview. *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)* 2006:52.
14. **Su Q,** Pavlov D, Chow JH, *et al*. Internet-Scale Collection of Human-Reviewed Data. *Proceedings of the 16th International Conference on World Wide Web* 2007:231—40.
15. **Snow R,** O'Connor B, Jurafsky D, *et al*. But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. *Proceedings of Empirical Methods in Natural Language Processing* 2008:254—63.
16. **Cohen J.** A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;**20**:37—46.
17. **Hripcsak G,** Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. *J Am Med Inform Assoc* 2005;**12**:296—8.
18. **Chinchor N,** Hirschman L, Lewis DD. Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3). *Computational Linguistics* 1993;**19**:409—49.
19. **Noreen EW.** Computer intensive methods for testing hypotheses: an introduction. New York: John Wiley & Sons, 1989.