Using Amazon's Mechanical Turk for Annotating Medical Named Entities

Meliha Yetisgen-Yildiz, PhD[1], Imre Solti, MD, PhD[1], Fei Xia, PhD[2]
[1]Biomedical & Health Informatics, School of Medicine, University of Washington, Seattle, WA
[2]Department of Linguistics, University of Washington, Seattle, WA

Abstract

Amazon's Mechanical Turk (AMT) service is becoming increasingly popular in Natural Language Processing (NLP) research. In this poster, we report our findings in using AMT to annotate biomedical text extracted from clinical trial descriptions with three entity types: medical condition, medication, and laboratory test. We also describe our observations on AMT workers' annotations.

Introduction

The manual construction of annotated corpora is extremely expensive both in terms of time and money. Snow et. al. (2008) demonstrated the potential power of Amazon's Mechanical Turk (AMT) service in creating large-scale annotated corpora for natural language tasks in a cheap and fast way[1]. We piloted the feasibility of using AMT for medical text annotation with 100 clinical trial announcements downloaded from ClinicalTrials.gov website.

Annotation Performance

To make the annotation task more convenient for AMT workers, we used a customized user interface and provided detailed annotation guidelines. Four AMT workers annotated the inclusion/exclusion sections of 100 selected announcements. We first posted the announcements to be annotated for medical condition, next for medication, and finally for laboratory test. We measured the quality of AMT annotations at different inter-annotator agreement levels by comparing the agreed entity spans to a gold standard (GS) manually created by one of the authors who has medical training. Agreement level k meant the annotation included only the spans that were agreed by at least k workers. As can be seen from Table 1, the annotation performance of non-medical expert AMT workers was very promising, especially for medical condition and medication.

Error Analysis

After AMT workers completed the tasks, we analyzed their annotations in detail in order to understand the problematic areas. This study led to the following observations for each entity type.

Medical Condition: As can be seen from Table 1, for agreement level k=1, the recall was almost perfect, R=0.99. On the other hand, the precision was lower, P=0.70 since some phrases (e.g., "*cardiac surgery*") annotated by the workers were not medical conditions. Such wrong annotations indicated that the workers were confused about the definition of medical condition.

k	Medical Condition			Medication			Laboratory Test		
	P	R	F	P	R	F	P	R	F
1	.70	**.99**	.79	.50	**.84**	.62	.42	**.73**	.53
2	.84	.87	**.86**	.79	.73	**.76**	.72	.65	**.68**
3	**.89**	.73	.80	**.93**	.45	.61	**.86**	.40	.54

Table 1. Quality measurement of AMT annotations. k: Agreement level, P: Precision, R: Recall, F: F-Measure.

Medication: For this task, the workers mainly failed to annotate many general GS medication phrases such as "*other investigational agents*" and collective names of groups of medications such as "*vitamins*". The existence of such errors indicated that either our guideline was not clear or descriptive enough for the workers or the workers did not pay enough attention to the guidelines.

***Laboratory Test*:** In clinical trials, the laboratory tests were usually represented as criteria with arithmetic comparator, such as "*hemoglobin level of >/=9.0 gm/dL*". The workers annotated almost all phrases with comparators (e.g. age>50) as test results which resulted very poor precision results for k=1.

For both medication and laboratory test, the workers wrongly annotated the other entity types. This might be a side effect of how we ordered the annotation tasks, since some workers (10 out of 72) worked on the annotation of multiple entity types. Those workers might have not read the guidelines for second and third tasks carefully because they thought they were annotating the previous entity type.

Conclusion and Future Plans

We believe that with careful design of the task AMT is a very promising tool for annotating biomedical text. For future work, we plan to improve the performance by revising annotation guidelines, increasing the number of annotations per announcement, and preventing the same workers from annotating different entity types.

Acknowledgment

This publication was made possible in part by: 1K99LM010227-0110.

References

1. Snow R, O'Connor B, Jurafsky D, Ng AY. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In Proceedings of EMNLP 2008.