Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements

Todd Lingren, Louise Deleger, Katalin Molnar, Haijun Zhai, Jareen Meinzen-Derr, Megan Kaiser, Laura Stoutenborough, Qi Li, Imre Solti

Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

Correspondence to
Dr Imre Solti, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, MLC 7024, Cincinnati, OH 45229-3039, USA; imre.solti@cchmc.org

IS is the senior author.

Received 26 March 2013
Revised 5 August 2013
Accepted 10 August 2013
Published Online First
3 September 2013

ABSTRACT

Objective To present a series of experiments: (1) to evaluate the impact of pre-annotation on the speed of manual annotation of clinical trial announcements; and (2) to test for potential bias, if pre-annotation is utilized.

Methods To build the gold standard, 1400 clinical trial announcements from the clinicaltrials.gov website were randomly selected and double annotated for diagnoses, signs, symptoms, Unified Medical Language System (UMLS) Concept Unique Identifiers, and SNOMED CT codes. We used two dictionary-based methods to pre-annotate the text. We evaluated the annotation time and potential bias through F-measures and ANOVA tests and implemented Bonferroni correction.

Results Time savings ranged from 13.85% to 21.5% per entity. Inter-annotator agreement (IAA) ranged from 93.4% to 95.5%. There was no statistically significant difference for IAA and annotator performance in pre-annotations.

Conclusions On every experiment pair, the annotator with the pre-annotated text needed less time to annotate than the annotator with non-labeled text. The time savings were statistically significant. Moreover, the pre-annotation did not reduce the IAA or annotator performance. Dictionary-based pre-annotation is a feasible and practical method to reduce the cost of annotation of clinical named entity recognition in the eligibility sections of clinical trial announcements without introducing bias in the annotation process.

OBJECTIVE

Natural language processing (NLP) projects require manually annotated gold standard corpora to train and test supervised, machine learning-based algorithms or, in the case of rule-based methods, to test the performance of the rules. In light of the high cost of expert manual annotations, NLP researchers need robust methods to speed up the annotation process, without biasing the generated gold standard. In our institution, we are working on an NIH-funded project to automate clinical trial eligibility screening by using NLP algorithms. This effort requires the development of a substantial manually annotated gold standard. As such, this annotation is very time-consuming and costly.

In this study, our aim is to present a series of experiments: (1) to evaluate the impact of pre-annotation on the speed of manual annotation of clinical trial announcements (CTA); and (2) to test for potential bias, if pre-annotation is utilized. We define potential bias as either increasing the discrepancy between annotators measured by inter-annotator agreement (IAA) or decreasing the agreement (called annotator performance in our study) between the annotations of the annotator with pre-annotated text and the eventual gold standard. The annotation task included labeling medical named entities in two classes: disease/disorder and sign/symptom. Unified Medical Language System (UMLS) Concept Unique Identifiers (CUI) and SNOMED-CT codes were also annotated for each entity.

The rest of the paper is structured as follows. In the 'Background and significance' section, we present relevant literature. In 'Data and methods', we describe the data, experimental methods, and analytical approaches. In the 'Results' section, we present the results. In the 'Discussion' section, we discuss the findings, limitations, and future research questions. In the final section, we provide our conclusions.

BACKGROUND AND SIGNIFICANCE

Pre-annotation has been studied widely in NLP tasks such as Named Entity Recognition (NER) (biomedical[1–4] and astrophysical[5] domains), part of speech (POS) tagging (Wall Street Journal[6–9] and medical literature[2 10]), and Semantic Frame/Role Labeling.[11] These approaches used some machine learning systems with varying sizes of training data. Some systems did active learning pre-annotation, incrementally training on iterative human input and presenting annotators with pre-annotated text,[2 5 8] while others[4 10] relied on an existing tool such as MetaMap[12] to generate a pre-annotation set to apply to the whole text.

Many applications for different domains have been built in order to semi-automatically annotate text as the user is working, updating future files with machine learning output based on previous annotations.[13–18] These efforts all seek to decrease annotation time, but in our study we focus on the role of a single pre-annotation set for particular named entities in the clinical domain. The main contribution of this study is in evaluating—in the clinical domain—if dictionary-based annotation sets provide substantial savings in time without biasing the annotation.

Several studies evaluated the time savings of pre-annotation. Using Wall Street Journal text, Ringger *et al*[6] studied the cost considerations of generating

To cite: Lingren T, Deleger L, Molnar K, *et al*. *J Am Med Inform Assoc* 2014;**21**:406–413.

a POS-tagged gold standard using many annotators and was able to reduce, by half, the amount of time it took to annotate the same amount of data. They concluded that the hourly cost savings were partially dependent on the (self-rated) expert level of the annotator.

In the biomedical domain, Ganchev *et al*[2] developed a semi-automated system to pre-annotate MEDLINE abstracts with a high-recall named entity tagger for gene mentions, and reported an astounding 75% reduction in time for the best tagger. In the clinical domain and using some of the same entity classes as our study, Ogren *et al*[4] used MetaMap to pre-annotate for disease and disorder. They reported a longer time for the pre-annotation set and doubted that there was any benefit in the pre-annotation method, citing spurious annotations that needed to be corrected. By building on these earlier works, we compare the performance and time savings from different annotation sets in the clinical domain. Our study is unique in the biomedical domain, as it evaluates the statistical significance of the potential bias effect of pre-annotation in addition to time and cost savings.

Machine learning-based pre-annotation is built upon training the model on a small amount of annotated text. A question remains whether the machine learning model is necessary in these cases, or whether a simple dictionary-based pre-annotation set is sufficient. Due to the initial smaller training set, the performance of a machine learning model is expected to be lower than a dictionary-based approach. We hypothesize that the dictionary-based approach might not have as many spurious results as Ogren *et al*'s approach; consequently, the dictionary-based pre-annotation will successfully reduce annotation time.

In evaluating the development of the Penn Treebank, Fort and Sagot[8] compared the quality of pre-annotation (using different POS taggers) and reported no significant difference in performance (Krippendorf's α[19]) between the two annotators, discounting that pre-annotation causes bias. Nor did Névéol *et al*[10] find bias from pre-annotation on semantic annotation of PubMed queries.

Other than in some limited domain set tasks, such as surname recognition[3] or POS tagging,[20] no dictionary-based pre-annotation method has been studied. Although not a dictionary method, pre-annotation of dates based on regular expressions was used to help decrease the time per annotation in a protected health information de-identification task of clinical notes.[21]

DATA AND METHODS

The annotation task in our study included annotating disease/disorder and sign/symptom entities. We followed the annotation guidelines and schema from the SHARPn project.[22] The SHARPn guidelines find and normalize clinically relevant mentions to Clinical Element Model templates, linking CUIs to mentions and identifying attributes and modifiers. We employed two experienced annotators (henceforth referred to as A1 and A2) with bachelor degrees who had been trained using these guidelines. One annotator had previous clinical expertise (as a registered nurse) and a Bachelor of Science degree in Nursing. Chapman *et al*[23] demonstrated that using both clinician and non-clinician annotators does not bias the annotated corpus, although non-clinicians need longer training time. The annotators were given access to the UMLS Terminology Services SNOMED CT[24] and Metathesaurus Browsers, in order to look up terms and assign CUIs and SNOMED-CT Codes (CODEs). The following is an example sentence from a CTA: 'Suspected of having lung cancer due to clinical symptoms, such as positive sputum cytology, hemoptysis, unresolved pneumonia, persistent cough…' A sample screen shot from the SNOMED-CT browser while searching for *lung cancer* is shown in figure 1.

Malignant tumor of lung is the best match for *lung cancer* and so the CODE (listed in the browser window as Concept: 363358000) and CUI (C0242379) are annotated with the span *lung cancer*. The five entities in the sample sentence (*lung*

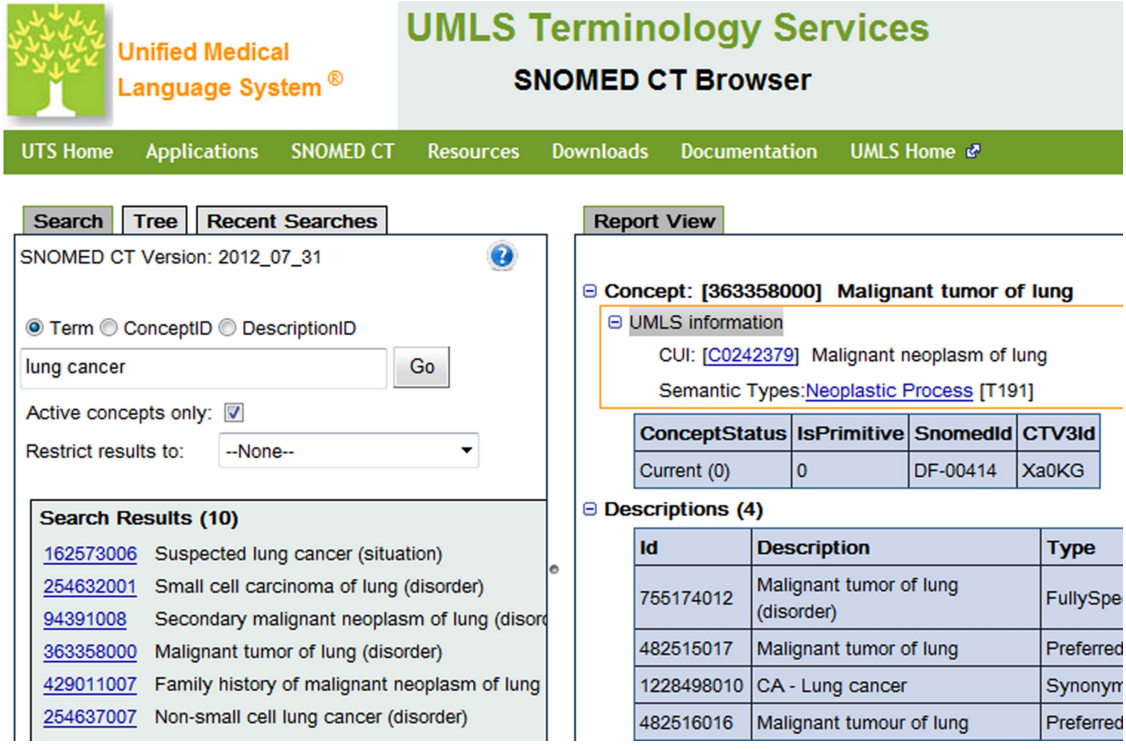

Figure 1 UMLS Technology Services SNOMED-CT browser: search for lung cancer.

cancer, *symptoms*, *hemoptysis*, *pneumonia*, and *persistent cough*) are all annotated with associated CUI and CODEs, as shown in figure 2.

Three entities belong to sign/symptom class and two are disease/disorder. Lab or test results (such as *positive x-ray* or *positive sputum cytology*) were not annotated. The Protégé plug-in Knowtator[25] was used for annotating the corpora. A screenshot from the program used to annotate is shown in figure 3.

Data

The CTA corpus for these experiments is composed of 1400 CTAs randomly selected from the clinicaltrials.gov website[26] (a total of 141 386 documents as of March 2013). We annotated only the eligibility criteria sections of the CTAs. One thousand of the 1400 CTAs were previously annotated[27] without pre-annotation for disease/disorder and sign/symptom. The 1000 were split in half and randomly assigned to a control group and a dictionary generation set. The control and dictionary generation sets are non-overlapping with the experiment sets. More detail is provided on the distribution of the remaining 400 CTAs into experiment sets in the 'Methods' section and in figure 4. The distribution of disease/disorder and sign/symptom entities for the CTAs was 196.3 tokens per file, with an average count of 7.1 entities per 100 tokens.

Methods

The two annotators were given sets of CTAs (both non-labeled and pre-annotated) to annotate in the Knowtator program for disease/disorder and sign/symptom. The sample size was determined based on the training size requirements of the Machine Learning algorithms that utilized the annotated CTAs. The underlying informatics projects provided the foundation for the exploratory pre-annotation experiments. The actual sample size is based not on the number of experiment CTAs (400) but on the units of analysis, namely the number of annotated entities and the number of tokens that the annotators read. Across all the control, dictionary, and experiment sets the annotators read almost 400 000 tokens (348 445) and annotated 19 002 medical named entities.

For the non-labeled text, annotators were asked to annotate disease/disorder and sign/symptom entities, as described above. For a pre-annotated text, annotators were given the following choices: removing an annotation they thought was spurious; keeping or modifying said annotation; or adding an additional annotation. Figure 3 depicts the Knowtator program, with a set of pre-annotations on a particular CTA for an annotator to remove, correct, approve, or add a new annotation. In adjudication all disagreements and any remaining ambiguities were resolved.

Pre-annotation procedure

Whereas previous studies relied on machine learning output to generate pre-annotation, we relied on a dictionary method in our study (figure 4). We evaluated two dictionaries of different sizes and origins, and each dictionary entry consisted of three items: term, UMLS CUI, and SNOMED-CT code. The first dictionary type was created by extracting annotations from the dictionary generation set of 500 CTAs, as described in the Data section. This dictionary is called the 'automated dictionary', as it represents the automatically extracted set of all of the annotations of the gold standard set. The CTA automated dictionary contains 3414 diseases/disorders and 294 signs/symptoms.

Entity Class	Span	CODE	CUI
DD [a]	*lung cancer*	363358000	C0242379
SS [b]	*symptoms*	404684003	C0037088
SS	*hemoptysis*	66857006	C0019079
DD	*Pneumonia*	233604007	C0032285
SS	*persistent cough*	284523002	C0562483

[a] disease/disorder [b] sign/symptom

Figure 2 Sample disease/disorder and sign/symptom entities.

The second dictionary type was created manually by the annotators, over several weeks. During the adjudication process of the double annotated, gold standard generation of the dictionary generation set of 500 CTAs, the annotators developed a list of what they determined to be common annotation decisions ('manual dictionary'). The CTA manual dictionary contains 522 disease/disorder entities and 47 signs/symptoms.

We used regular expression matching to pre-annotate the text, with the dictionary terms as input (see figure 4, Experiments). The list of matches and their offsets was imported into Knowtator in order to assign the class labels for each term. We wrote a program to assign the UMLS CUIs and SNOMED-CT codes to the pre-annotated terms. Table 1 shows the number of dictionary matches for each experiment set.

We split the text for each experiment into two sets, Set1 and Set2. A1 received non-labeled text in Set1 and pre-annotated text in Set2; A2 received pre-annotated text in Set1 and non-labeled text in Set2. Table 1 details each of these sets, as follows: the total number of entities for each set; the annotator who had the pre-annotated set (A1 or A2); the dictionary type that was used for the pre-annotation (manual vs automated); and the hypothesis tested in the experiment. For the dictionary and control sets, the number of entities shown is the number of entities in the gold standard. For the experiment sets, the number is the result of pre-annotation (the number of entities given to the annotator with pre-annotation). Figure 4 also details the study design for the experiments.

Experiments

There are two experiments (labeled: 1, 2 in table 1). As shown in figure 4, each experiment is split into two sub-experiment sets (eg, 1.1, 1.2; 2.1, 2.2). The first document set (listed as 'Dictionary') includes 500 traditionally-annotated, gold standard CTAs and is the source of pre-annotation terms for experiments 1 and 2. The second document set (listed as 'Control') includes 500 traditionally-annotated, gold standard CTAs. The experiment sets 1 and 2 comprise the remaining 400 CTAs for experiments. Each experiment set was double annotated and adjudicated for a final gold standard.

Experiment 1

A1 was given 100 non-labeled CTAs in set 1.1 and 100 pre-annotated CTAs in set 1.2. A2 was given pre-annotated CTAs in set 1.1 and non-labeled CTAs in set 1.2. The purpose of this experiment was to evaluate the potential bias of the *CTA manual dictionary pre-annotation* on the annotator and potential pre-annotation time savings using terms for pre-annotation that were collected by the annotators in their earlier CTA annotation projects.

Experiment 2

The purpose of this experiment was to evaluate the potential bias of the *CTA automated dictionary pre-annotation* on the annotator and potential pre-annotation time savings. A1 was given 100 non-labeled CTAs in set 2.1 and 100 pre-annotated

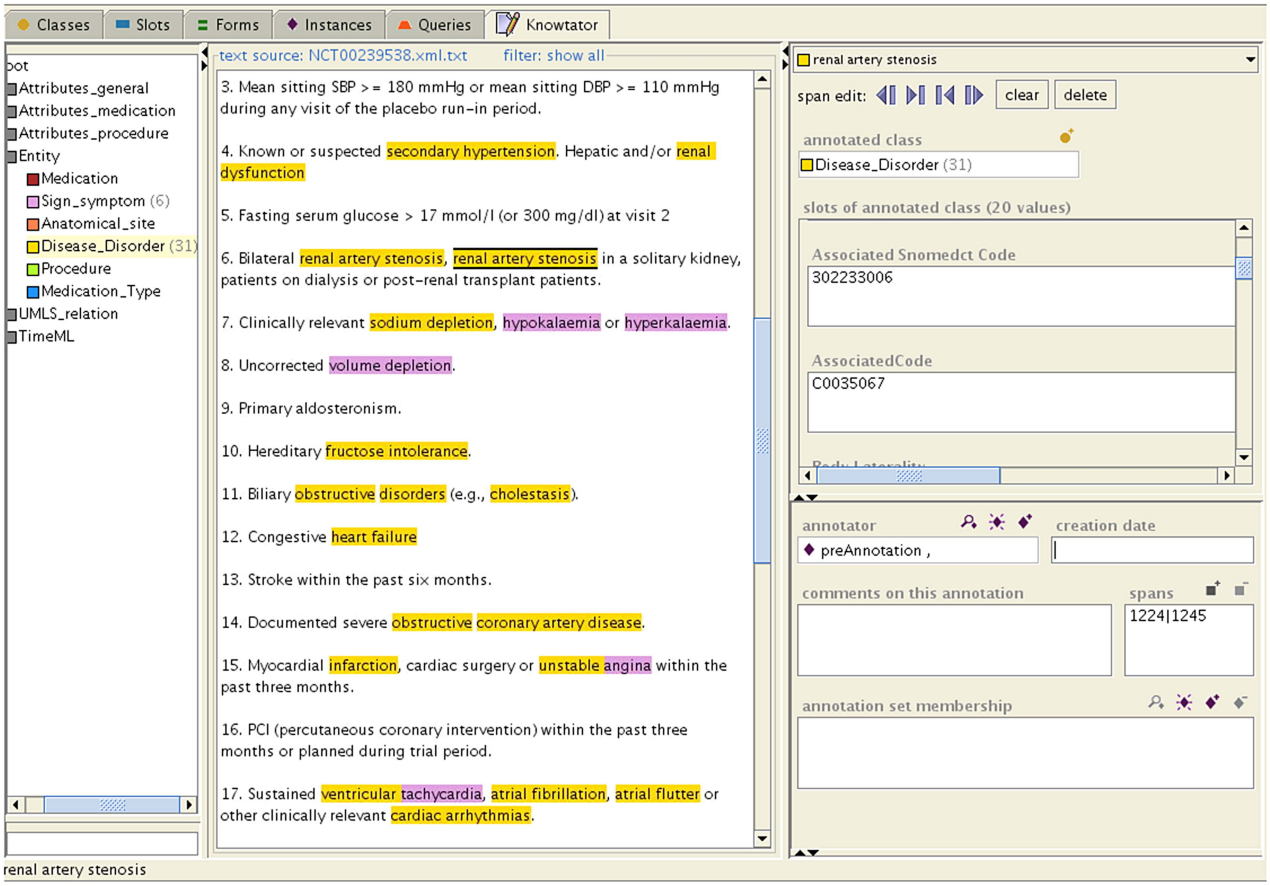

Figure 3 Pre-annotated clinical trial announcement text in Knowtator.

CTAs in set 2.2. A2 was given pre-annotated CTAs in set 2.1 and non-labeled CTAs in set 2.2. The purpose of doing both experiments 1 and 2 is to compare how different pre-annotation dictionary types (automated and manual in the CTA corpus) affects the IAA, performance relative to the eventual gold standard (resulting from the adjudication process), and potential time savings.

RESULTS

Measuring annotator bias

By comparing the IAA for each set in an experiment (eg, experiment 1: sets 1.1 and 1.2), we looked for potential bias caused by annotating text with pre-annotation. The F-measure (equation 3) calculated is the harmonic mean between precision (equation 1) and recall (equation 2).

$$P = \text{True Positives}/(\text{True Positives} + \text{False Positives}) \quad (1)$$

$$R = \text{True Positives}/(\text{True Positives} + \text{False Negatives}) \quad (2)$$

$$F = 2 \times P \times R/(P + R) \quad (3)$$

The IAA compares the agreement between each annotator by temporarily treating one annotator (eg, A1) as the gold standard and calculating the F-measure for the other annotator (eg, A2).[28] When we report on the F-measure IAA, we list only one per class because the F-measure is identical for each annotator (A1's precision relative to A2 is A2's recall relative to A1).

Measuring individual annotators' distance from adjudicated gold standard

After the double annotation of each experiment set, the annotators met in adjudication (under the supervision of one of the investigators) and came to an agreement on a final gold standard. An F-measure was calculated for each annotator, relative to the gold standard for each entity class (disease/disorder and sign/symptom) for that set. This is what we are calling the annotator's *performance*.

Comparing the performance between the annotator who received non-labeled text and the annotator who received pre-annotated text, within a single experiment set (eg, 1.1), helps to show any potential biasing effect that pre-annotation has on the annotators' performances, relative to the gold standard. We can also compare the same annotator's (eg, A1) annotation speed, in the experiment set with non-labeled text (eg, 1.1), with the experiment set with pre-annotated text (1.2).

The impact of pre-annotation on annotation speed is measured for the same annotator across sub-experiments (eg, 1.1 vs 1.2), using the same corpus and dictionary approach, while the impact of pre-annotation on creating bias is measured between annotators within sub-experiments (eg, A1 vs A2 in 1.1 and then again in 1.2). The experiments are repeated two times, for dictionary differences (manual vs automated). In addition, there is a further multiplying factor of two based on which annotator is getting pre-annotated text. Altogether there are four sub-experiments (as shown in table 1) to control for dictionary and pre-annotation.

Statistical analysis

We performed one-way analysis of variance (ANOVA) on nine variables: A1 F-measure against the gold standard, for disease/

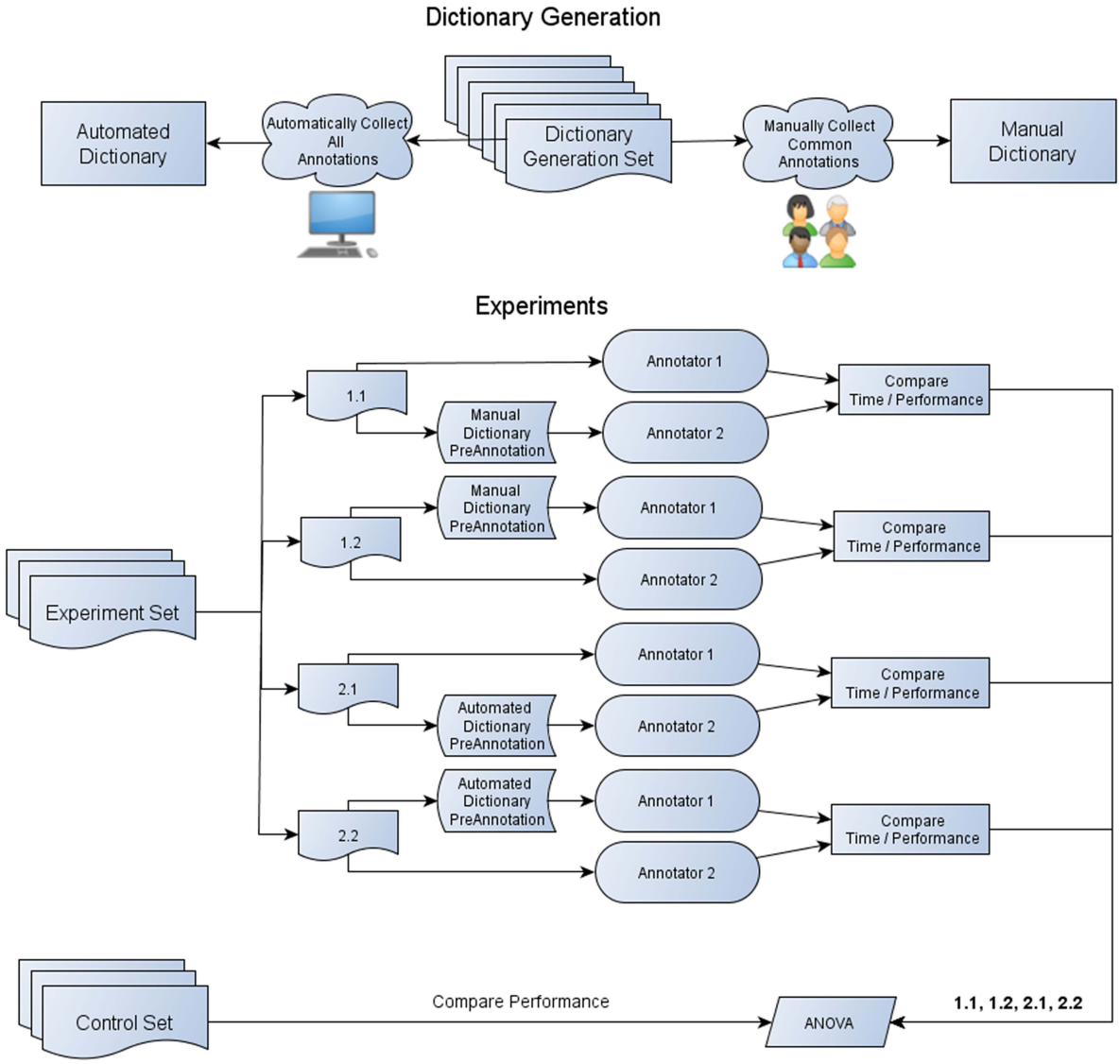

Figure 4 Experiment study design.

Table 1 CTA pre-annotation experiments

Document sets	Corpus	Number of files	Entity class DD	Entity class SS	Annotator with pre-annotated Text	Dictionary method	Hypothesis
Dictionary	CTA	500	6478	484	N/A	N/A	
Control	CTA	500	8117	474	N/A	N/A	
Experiment							
1							
1.1	CTA	100	719	39	A2	Manually generated	Using human annotator collected dictionary of annotation terms to pre-annotate CTAs will reduce annotation time without accompanied bias
1.2	CTA	100	603	38	A1		
2							
2.1	CTA	100	878	102	A2	Automatically generated	Using automatically generated dictionary of annotation terms to pre-annotate CTAs will reduce annotation time without accompanied bias
2.2	CTA	100	994	76	A1		

A1, annotator 1; A2, annotator 2; CTA, clinical trial announcements; DD, disease/disorder; SS, sign/symptom.

Table 2 IAA and annotator performance

Experiment set	IAA (%)	Performance (%)	
		A1	A2
1.1	95.5	98.8	96.4
1.2	93.4	98.2	95.2
2.1	93.7	97.0	96.0
2.2	94.7	97.0	96.9

A1, annotator 1; A2, annotator 2; IAA, inter-annotator agreement.

disorder annotation; A2 F-measure against the gold standard, for disease/disorder annotation; A1 F-measure against the gold standard, for sign/symptom annotation; A2 F-measure against the gold standard, for sign/symptom annotation; A1 versus A2 IAA, for disease/disorder annotation; A1 versus A2 IAA, for sign/symptom annotation; number of class entities, tokens, and CUI/CODE entities.

The purpose of performing an ANOVA test on each of these variables was to determine if the variance between files and annotators was statistically significant. Due to the number of different tests conducted, we applied a very conservative Bonferroni correction to account for the increased possibility of type I error. Thus, to adjust for nine different significance tests with multiple variables that may not be independent,[29] findings were considered statistically significant at $p<0.0001$.

The sets for comparison were the control documents for each experiment set, both as pairs (1.1/1.2, etc.) and individually (1.1 vs Control, 1.2 vs Control, etc.).

To calculate statistical significance, in order to test whether the IAA between annotators or whether each annotator's performance was significantly different among experiment sets, we calculated the F-measures per files. These per-file F-measures were compared using a one-way ANOVA test for statistical significance among the different experimental groups. Table 2 shows the per-set averages of the F-measures for IAA and annotator performance.

Time savings

To test for time savings in annotation for each set, we recorded the annotation times and compared them to evaluate the effect of pre-annotation.

Table 3 displays the time savings for pre-annotated text over non-labeled text. For each experiment set, the amount of time needed to annotate and calculated time savings of annotating with pre-annotated text are indicated. Also included is the average of both sets of each experiment. For example, set 1.1 took 17.7 h for pre-annotated text and 20.5 h for non-labeled text. This represents an overall time savings of 13.9% for A2, who had pre-annotated text. Also in set 1.1, A2 took an average of 45.4 s per entity with pre-annotated text, while A1 took an average of 7.3 s longer per entity with non-labeled text. The average between the two sub-experiment 1 sets was 16.6% for overall time and for per-entity time savings. The greatest overall time savings is in set 2.2 (automated dictionary pre-annotation) with 20.8%. A paired t test shows that the time savings in each experiment set were significant ($p<0.01$).

Comparisons for statistical significance

Each experiment set has three F-measures (agreement between the annotators and performance for each annotator). The performance reported is the combined class F-value, which is the F-measure for both classes of disease/disorder and sign/symptom; these are listed in table 2. Table 4 lists the p values from the results of the ANOVA comparisons for each experiment pair. The purpose of this comparison is to examine if there is a significant difference in an annotator's performance when receiving pre-annotated or non-labeled text. The annotator F-measures are separated according to entity class. The control 500 CTA set, where no pre-annotation occurred, provides a set for comparison. In each column, an experiment set is compared against the control set. In the first column the pooled CTA text sets (1.1–2.2) were compared against the control set. In the second column, the set 1.1 was compared, and so on.

The results in table 4 show that when annotating signs and symptoms, the annotators' performance and IAA are significantly different from the eventual gold standard on Bonferroni $p<0.0001$ level. This finding is significant for experiment sets 1.1 and 1.2. None of the other comparisons show statistically significant difference.

Table 4 also lists the p value for each variable in intra-experiment ANOVA comparisons. There is no statistically significant difference between manual and automated experiments.

DISCUSSION

Time savings

In every experiment pair, the annotator with the pre-annotated text took less time to annotate than the annotator with non-labeled text. This illustrates a clear time savings and, unlike other studies,[4] spurious annotations in the pre-annotation set did not affect the annotator's performance. The time saved in each experiment was significant ($p<0.01$). The time savings result in part from reducing the amount of time an annotator has to look up entities to match with the UMLS terminology database (see figure 1).

The automated dictionary pre-annotation experiment (2.1/2.2) shows greater per-entity time savings, compared to the manual dictionary experiment sets (20.8% time savings vs 16.6%). The reduced time savings of the manual dictionary-based pre-annotation set versus the automated may be due to a lack of coverage, since the automated dictionary contained more than six times the total entries (3708 vs 569). During adjudication we learned that many of the smaller abbreviations (eg, 'ms' (multiple

Table 3 Overall and per entity time savings

Experiment set	Overall time (hours)				Time per entity (seconds)				p Value
	Pre-annotated text	Non-label	% Saved	Average per experiment (%)	Pre	Non-label	% Saved	Average per experiment (%)	
1.1	17.7	20.5	13.9		45.4	52.7	13.9		<0.01
1.2	14.3	17.7	19.3	16.6	34.9	43.3	19.3	16.6	<0.01
2.1	14	17.5	20.0		30.5	38.2	20.0		<0.01
2.2	14.25	18.2	21.5	20.8	28.7	36.6	21.5	20.8	<0.01

Table 4 Statistical significance of experiments

Statistical significance of experiments 1–2 (CTA)					
	CTA vs 500*	1.1 vs 500	1.2 vs 500	2.1 vs 500	2.2 vs 500
A1 vs GS (D)	0.37	0.38	0.43	0.44	0.08
A1 vs GS (S)	0.42	**<0.0001**	**<0.0001**	0.3	0.98
A2 vs GS (D)	0.36	0.14	0.01	0.22	0.11
A2 vs GS (S)	0.38	**<0.0001**	**<0.0001**	0.57	0.01
IAA (D)	0.34	0.96	0.24	0.54	0.95
IAA (S)	0.06	**<0.0001**	**<0.0001**	0.38	0.73
Code_Ent	0.35	0.14	0.28	0.48	0.22
DS_Ent	0.45	0.13	0.03	0.98	0.16
Tokens	0.4	0.06	0.03	0.11	0.15

Intra-experiment significance		
	1.1 vs 1.2	2.1 vs 2.2
A1 vs GS (D)	0.2	0.06
A1 vs GS (S)	0.47	0.76
A2 vs GS (D)	0.45	0.83
A2 vs GS (S)	0.43	0.57
IAA (D)	0.12	0.48
IAA (S)	0.29	0.46

*Control for CTA.
A1, annotator 1; A2, annotator 2; CTA, clinical trial announcements;
D, disease/disorder; GS, gold standard; IAA, inter-annotator agreement; S, sign/symptom.
Bold indicates statistical significance at p<0.0001.

sclerosis), 'all' (acute lymphoblastic lymphoma)) produce spurious annotations that cost time in removing. The pre-annotation program performed the lookup and annotated without regard to capitalization, matching complete tokens only. For example, the abbreviation MS would match both 'ms' and 'MS', but not 'aims'. Modifications to the pre-annotation program could be developed to allow shorter (two to three letter) abbreviations to be case sensitive and further increase time savings for pre-annotated tasks.

To put the time savings in perspective, a 3-month (60 work days) annotation project can be reduced to as little as 48 days when using an automatically generated pre-annotation dictionary on CTAs. When using a manually created pre-annotation dictionary on the same corpus, the 60 days can be reduced to 50. For projects that implement double annotation, the saved labor cost is twice of the saved time, as 10 days annotation time saves 20 days' labor.

Performance

Compared to the eventual gold standard, the annotator without pre-annotation missed short abbreviations more often, including 'v' (vomiting), 'uti' (urinary tract infection), and 'mm' (multiple myeloma). Pre-annotation can capture these short tokens. However, the annotator without pre-annotation missed fewer long phrasal annotations which require close reading of the text such as 'lack of progress in his speech sound development' and 'decreased active rotation range of motion'. In addition, although the time savings were significant, the annotator with pre-annotated text tended to allow frequently occurring terms like 'disease' or 'infections' to remain unmodified, even if there were additional qualifying terms like 'autoimmune' or 'hepatitis'.

Comparisons for statistical significance

The purpose of performing an ANOVA test on each of the nine variables was to determine if any of the variance was statistically significant. We demonstrated that the class entities, tokens, and CUI/CODE entities were not statistically significantly different, in most of the set comparisons, when compared to the baseline set. This indicates that the texts' structures are not so different as to cause annotation speed differences. In CTAs, sign/symptom entities are not as frequent as disease/disorder and only average 0.9 entities per file, or 0.32 per 100 tokens. We believe that rare sign/symptoms entities in this corpus did not provide a strong basis for the statistical significance test and this is the reason why the sign/symptom IAA and annotator performance were statistically significantly different in experiment 1 (table 4).

Another important comparison point is the intra-experiment ANOVA calculation for annotator performances. This indicates the potential statistical significance of the variance between the annotator who had pre-annotation and the annotator who did not, between manual and automated dictionary experiments. In no category of F-measure was the performance difference statistically significant to a Bonferroni corrected p value of 0.0001 (table 4, intra-experiment significance); that is, the pre-annotation did not introduce annotation bias.

Limitations

A limitation of this study is that annotation time savings and potential annotation bias are not tested in the same sub-experiments. However, this is mitigated by the careful study design and ANOVA tests. Another limitation is the focus on just one corpora (CTA) and one source of dictionaries. Although preliminary results showed a similar pattern for pre-annotation experiments on a clinical note corpus, further research is needed with multiple different clinical corpora. Future studies should also experiment with other dictionary sources such as UMLS. Finally, cross-corpora pre-annotation experiments have been planned with dictionaries generated from different types of clinical texts. For practical purposes it is a limitation of our proposed method that we did not test the pre-annotation value of the dictionaries based on the number of underlying documents. That is, we used a fixed set of 500 documents to generate the dictionaries instead of consecutively increasing sets (eg, 100, 200, and so on documents). Future research should test if a dictionary based on smaller number of notes would have a beneficial effect.

CONCLUSIONS

This study evaluated the effects of pre-annotation on annotation time and annotator bias in the annotation of disease/disorder and sign/symptom entities for an important clinical corpora, CTAs. The pre-annotated set was created from either an automatically extracted or a manually generated dictionary. Time savings were statistically significant and present in all of the experiments, when the annotator used pre-annotated text. There was no statistically significant difference in annotator performance or IAA between using a manually or automatically collected dictionary of pre-annotation sets. Furthermore, pre-annotated text did not introduce bias for the annotations. We conclude that either manually or automatically generated dictionary-based pre-annotation is a feasible and practical method to reduce the cost of clinical NER in the eligibility sections of CTAs without introducing bias in the annotation process.

Contributors TL supervised the annotations, ran 80% of the experiments, and wrote the first draft of the manuscript. LD ran 10% of the experiments and contributed to the annotation guidelines. KM wrote the code matching pre-annotation program and contributed ideas for experiments. HZ wrote a regular expression pre-annotation program and ran 5% of the experiments. JM-D contributed ideas for statistical analysis. QL contributed ideas for algorithm development and ran 5% of the experiments. MK and LS annotated the NLP corpus. This manuscript was prepared by TL and IS with additional contributions by all authors. IS and TL designed the experiments and IS supervised the project.

Funding The authors and annotators were supported by internal funds from Cincinnati Children's Hospital Medical Center. IS, LD, TL, HZ, and LS were partially supported by grants 5R00LM010227-04, 1R21HD072883-01, and 1U01HG006828-01.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Tomanek K, Wermter J, Hahn U. Efficient annotation with the jena annotation environment (JANE). *Linguistic Annotation Workshop—A Merger of NLPXML 2007 and FLAC 2007*; Prague, Czech Republic: Association for Computational Linguistics (ACL), 2007:9–16.
2 Ganchev K, Pereira F, Mandel M, *et al*. Semi-automated named entity annotation. *Proceedings of the Linguistic Annotation Workshop*; Prague, Czech Republic: Association for Computational Linguistics, 2007:53–6.
3 Aramaki E, Miura Y, Tonoike M, *et al*. Extraction of adverse drug effects from clinical records. *Stud Health Technol Inform* 2010;160(Pt 1):739–43.
4 Ogren P, Savova G, Chute C. Constructing evaluation corpora for automated clinical named entity recognition. *Proceedings of the Language Resources and Evaluation Conference (LREC)*; 2008:28–30.
5 Hachey B, Alex B, Becker M. Investigating the effects of selective sampling on the annotation task. *Proceedings of the Ninth Conference on Computational Natural Language Learning*; Ann Arbor, Michigan: Association for Computational Linguistics, 2005:144–51.
6 Ringger E, Carmen M, Haertel R, *et al*. Assessing the costs of machine-assisted corpus annotation through a user study. *Proceedings of the Language Resources and Evaluation Conference (LREC)*; 2008.
7 Marcus MP, Marcinkiewicz MA, Santorini B. Building a large annotated corpus of English: the Penn Treebank. *Comput Linguist* 1993;19:313–30.
8 Fort K, Sagot B. Influence of pre-annotation on POS-tagged corpus development. *Proceedings of the Fourth Linguistic Annotation Workshop*; Uppsala, Sweden: Association for Computational Linguistics, 2010:56–63.
9 Chiou FD, Chiang D, Palmer M. Facilitating treebank annotation using a statistical parser. *Proceedings of the first international conference on human language technology research*; San Diego: Association for Computational Linguistics, 2001:1–4.
10 Névéol A, Islamaj Doğan R, Lu Z. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. *J Biomed Inform* 2011;44:310–18.
11 Jiang J, Zhai CX. A two-stage approach to domain adaptation for statistical classifiers. *Proceedings of the Sixteenth Conference on Information and Knowledge Management*; Lisbon, Portugal: ACM, 2007:401–10.
12 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium: American Medical Informatics Association*; 2001:17.
13 Salgado D, Krallinger M, Depaule M, *et al*. MyMiner: a web application for computer-assisted biocuration and text annotation. *Bioinformatics* 2012;28:2285–7.
14 Felt P, Merkling O, Carmen M, *et al*. CCASH: a web application framework for efficient distributed language resource development. *Proceedings of Language Resources and Evaluation Conference (LREC)*; 2010.
15 South BR, Shen S, Leng J, *et al*. A prototype tool set to support machine-assisted annotation. *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP 2012)*; 8 June 2012, 2012:130–9.
16 Fragkou P, Petasis G, Theodorakos A, *et al*. BOEMIE ontology-based text annotation tool. *Proceedings of the Language Resources and Evaluation Conference (LREC)*; 2008:28–30.
17 Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* 2005;21:3191–2.
18 Stenetorp P, Pyysalo S, Topic G, *et al*. BRAT: a web-based tool for NLP-assisted text annotation. *EACL* 2012;2012:102.
19 Krippendorff K. *Content analysis: an introduction to its methodology*. Sage Publications, 2004.
20 Carmen M, Felt P, Haertel R, *et al*. Tag dictionaries accelerate manual annotation. *Proceedings of Language Resources and Evaluation Conference (LREC)*; 2010.
21 South BR, Shen S, Friedlin FJ, *et al*. Enhancing annotation of clinical text using pre-annotation of common PHI. *Proceedings of the AMIA Annual Symposium*; 2010:1267.
22 Albright D, Lanfranchi A, Fredriksen A, *et al*. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *J Am Med Inform Assoc* 2013; 20:922–30.
23 Chapman WW, Dowling JN, Hripcsak G. Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. *Int J Med Inform* 2008;77:107–13.
24 UMLS Terminology Services. https://uts.nlm.nih.gov/home.html (accessed 23 Mar 2013).
25 Ogren PV. Knowtator: a protégé plug-in for annotated corpus construction. In: *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*; 2006:273–5. http://knowtator.sourceforge.net/
26 http://www.clinicaltrials.gov/ct2/home/
27 Li Q, Zhai H, Deleger L, *et al*. A sequence labeling approach to link medications and their attributes in clinical notes and clinical trial announcements for information extraction. *J Am Med Inform Assoc* 2013;20:915–21.
28 Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. *J Am Med Inform Assoc* 2005;12:296–8.
29 Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ* 1995;310:170.