MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition

Philip V. Ogren, Guergana K. Savova , Christopher G. Chute

Division of Biomedical Informatics, Mayo Clinic College of Medicine, USA

Abstract

We report on the construction of a gold-standard data set consisting of annotated clinical notes suitable for evaluating our biomedical named entity recognition system. The data set is the result of consensus between four human annotators and contains 1,556 annotations on 160 clinical notes using 658 unique concept codes from SNOMED-CT corresponding to human disorders. Inter-annotator agreement was calculated on annotations from 100 of the documents for span agreement (91% as positive specific agreement). For annotations found to be a match with respect to span, agreement was calculated for concept code (82% as positive specific agreement), context (85% as Kappa), and status (89% as Kappa). Complete agreement for span, concept code, context, and status was 75% as positive specific agreement. We consider this to be a conservative estimate for the upper bound for total system performance evaluated with the gold-standard. We found that creating a consensus set based on annotations from two independently created annotation sets can reduce inter-annotator disagreement by 32.3%. Surprisingly, we found little benefit to pre-annotating the corpus with a third-party named entity recognizer.

Keywords:

Natural language processing, named entity recognition, annotated corpora, inter-annotator agreement, SNOMED-CT

Introduction

The text analysis group at the Mayo Clinic College of Medicine has built and deployed a text analysis system that processes the entire repository of clinical notes in Mayo's electronic medical record. The system identifies, among other things, mentions of disorders and the context and status of those mentions. This output provides a concept-based index of the clinical notes repository used for retrieval purposes by researchers. We have anecdotal evidence based on browsing system output and end-user feedback that gives us a rough idea of the quality of our system. However, there has not yet been a formal evaluation of the final output of the system. By creating a gold-standard data set that represents complete and accurate system output, it becomes possible to measure the performance of actual system output in a way that is automated and comparable across multiple versions of the system. To this end, we have created a gold-standard corpus of annotated clinical notes for the disorder concepts that our system attempts to identify. In this paper we describe how we constructed this gold-standard corpus and evaluated its quality.

[1] provide a survey of gold-standard corpora that are available for the biomedical domain. They are few in number and vary widely in size and quality and none of them were designed with clinical data in mind. There have been no such publicly-available resources for the clinical domain until very recently[1]. This is due to the sensitive nature of patient data and the strict confidentiality laws designed to protect them[2]. Unfortunately, the development of gold-standard corpora is difficult and expensive because of the tedious and detailed nature of the work and the domain expertise required. Some of the well known challenges are outlined in [2]. Absent shared community resources for the medical domain, we are faced with the choice of creating our own or doing without. As we took up the challenge to build our own gold-standard data set, a key consideration was to determine the feasibility of creating a very high quality corpus with limited resources. Thus, the process of building our gold standard was designed so that we could measure the expense of additional review of the annotations against gains in quality of the data set. We also explored one way to potentially speed up annotation: pre-annotation with a third party system, MetaMap.

Materials and methods

Clinical notes corpus

The Mayo Clinic has a repository of approximately twenty million clinical notes that consist of documents dictated by physicians that are subsequently transcribed and filed as part of the patient's electronic medical record. The notes consist of sections such as *Chief Complaint, Current Medications,* and *Impression/Plan,* among others. The repository contains outpatient notes, discharge summaries, and inpatient service notes. From this repository we ran-

1 See https://www.i2b2.org/NLP/.

2 For this reason, our corpus will not be made publicly available in its current form. Contact chute@mayo.edu further information.

domly selected 160 notes for the corpus used for the gold-standard data set. The total number of words in the corpus is 47,975 with a median word count of 249 words per note.

Table 1 - UMLS Semantic types corresponding to disorders

TUI	type name
T019	Congenital abnormality
T020	Acquired abnormality
T037	Injury or Poisoning
T046	Pathologic Function
T047	Disease or Syndrome
T048	Mental or Behavioral Dysfunction
T049	Cell or Molecular Dysfunction
T050	Experimental Model of Disease
T190	Anatomical Abnormality
T191	Neoplastic Process

Annotation schema

The annotation task performed by the annotators consists of creating labeled spans of text that correspond to mentions of disorders found in the clinical note. Each annotation has one or more spans of text, a concept code, a context, a status, and a flag that indicates whether the mentioned disorder is related to the patient or not. An annotation's concept code is a string attribute that contains a concept identifier from a controlled vocabulary, in this case SNOMED-CT. The context of an annotation is an attribute with one of the following values: *current, history of, and family history of.* The status of an annotation is an attribute with one of the following values: *confirmed, possible,* and *negated*. Any combination of context and status values is permitted for an annotation. For example, the combination of the context *family history of* with the status *negated* means that the patient has no family history of the mentioned disorder. An annotation may also be flagged as unrelated to the patient if the mentioned disorder has little or nothing to do with the patient's health (e.g. an organization's name as in "Diabetes Clinic"). When an annotation is flagged as unrelated to the patient, the context and status are both automatically given the value *unrelated to patient.*

We created a subset of SNOMED-CT that contains only those concepts corresponding to disorders by leveraging the Unified Medical Language System[3] (UMLS) and its Semantic Network. The UMLS assigns to each concept in SNOMED-CT one or more semantic types defined in the Semantic Network. The subset of SNOMED-CT we used was created by selecting only those SNOMED-CT concepts assigned one of the semantic types shown in Table 1. This list was derived from [3][4]. The resulting subset of SNOMED-CT consists of 82,813 concepts and was provided to the annotators via an interface that provides keyword search and hierarchical navigation[5].

Annotators

Four clinical data retrieval experts performed the annotation task. All of them have been in their current work positions since 2002 and have prior experience in the medical coding of patient records. Each of the annotators has experience with various medical terminologies (e.g. ICD-9). However, none of them had prior experience with SNOMED-CT and no SNOMED-CT specific training was conducted.

Annotation software

Knowtato[6]r is a general purpose text annotation tool for creating gold-standard training and evaluation corpora for Natural Language Processing (NLP) systems described in [4]. The annotation schema described above was easily instantiated in Knowtator without any task-specific software development. Knowtator provides a mechanism to aggregate multiple sets of annotations into a single annotation project and various inter-annotator agreement (IAA) metrics for comparing multiple sets of annotations. Additionally, Knowtator provides a consensus set creation feature that provides a mechanism for resolving differences between two sets of annotations.

Table 2 - Annotation summary

set	description	count	hours
A1	MMTx pre-annotated	1105	53.5
A2	not pre-annotated	1142	53
A3	MMTx pre-annotated	1054	38
A4	not pre-annotated	1113	21
C1	consensus of A1 & A3	1125	8
C2	consensus of A2 & A4	1193	11
GS	consensus of C1 & C2	1164	13
MMTx	created by MMTx	664	<1

3 http://www.nlm.nih.gov/pubs/factsheets/umls.html. ver. 2005AC

4 We excluded *Finding* and *Signs and Symptoms* to limit scope.

5 See http://www.nlm.nih.gov/research/umls/meta6.html#s6_11

6 See http://bionlp.sourceforge.net/Knowtator

Pre-annotation of the corpus with MMTx:

We pre-annotated all 160 notes in the corpus using the MetaMap Transfer (MMTx)[7] tool. MMTx provides mappings from natural language text to UMLS concepts. The MMTx distribution provides a set of scripts that can build a customized target set of concepts to map to. Using these scripts we were able to provide MMTx with the set of SNOMED-CT concepts described in §2.1. Each annotation that MMTx presents has a span, a concept code, and a relevance score. We used only those mappings with a relevance score of 900 or greater (out of a possible 1000) in order to reduce the number of spurious matches. The resulting annotations were imported into Knowtator and were made available to two of the annotators as described below.

Annotation workflow

Each of the 160 notes was annotated individually by all four annotators, referred to as A1, A2, A3, and A4 (see Table 2). Two of the annotators, A1 and A3, were given MMTx annotations and were encouraged to use, discard or add to the provided annotations as they saw fit. The other two annotators, A2 and A4, were given the clinical notes without any MMTx annotations provided. Other than the presence or absence of MMTx annotations, all four annotators had exactly the same annotation task using the same annotation guidelines. After a note was annotated individually by A1 and A3, a set of pair-wise consensus annotations, C1, was created by A1 and A3 using the annotations they created individually and having them work together to resolve all differences. Similarly, a consensus set, C2, was created from A2 and A4's annotations. When both consensus sets were created for a document a master consensus set, GS, was created by all four annotators working together to resolve the differences between C1 and C2. The set GS is considered our final gold-standard set of annotations.

Sixty of the notes were annotated during a trial phase during which the annotation schema was finalized, the annotation guidelines were completed, and the annotators were trained. Frequent meetings were held during this phase to provide instruction, review examples, and receive feedback about the task. IAA results reported here exclude annotations from this set of 60 notes. However, we included these annotations into our final gold-standard data set because the annotation schema and guidelines were completed before a final round of four-way consensus annotation was performed as described above. During the annotation of the remaining 100 notes the annotation schema and guidelines were not changed, no meetings were held, and the annotators were given a strict prohibition from communicating with each other about the annotation task. IAA results reported below are only for these 100 notes.

Annotation guidelines

[5] show that careful consideration of annotation guidelines can have a marked impact on IAA. Therefore, we took care to create detailed and complete guidelines for the annotators. The annotation guidelines given to the annotators consist of about 3900 words and contain over 40 examples. The guidelines have the following four overarching principles:

1. *A mentioned disorder should be assigned the most specific concept code named by the span of text*. The text covered by the span of an annotation should be a reasonable synonym of the names associated with the concept. Descriptions of disorders should not be annotated, i.e. there should be little or no inference performed by the annotator.
2. *All mentions of disorders in each note are to be annotated.* Every disorder mentioned by name in the text of the clinical note should be annotated regardless of its relevance to the patient. If the disorder is not related to the patient's health, then the corresponding annotation should be flagged as *unrelated to patient*.
3. *A disorder is defined as any concept that appears in the subset of SNOMED-CT that has been provided.* This rule provides an unambiguous and clear definition of what a disorder is.
4. *There should be only one annotation per mentioned disorder.* In some cases a mentioned disorder could reasonably be assigned to more than one concept code. Always, choose the most specific concept that is named by the text.

The remainder of the guidelines consist of a detailed description of the annotation schema, Q&A styled instructions, and examples. Some example questions answered in the guidelines were: *How do I decide what the most specific concept is? Are nested annotations allowed? Are overlapping annotations allowed? Can the spanned text of an annotation be more specific than the assigned concept code?*

Calculating IAA

We report IAA as positive specific agreement described in [6] given by:

$$Ppos = \frac{2a}{2a + b + c} \tag{1}$$

7 See http://mmtx.nlm.nih.gov/

where a is the number of cases where the annotators agree and where b and c are non-matches from each annotator, respectively. Positive specific agreement is reported for match criteria requiring span selection agreement because negative instances are not easily counted and expected chance agreement is vanishingly small. We also report the Kappa statistic described by [7] and [8] given by:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{2}$$

where P(A) is percentage agreement and P(E) is expected chance agreement. Kappa is reported for match criteria that compare only annotations that already agree with respect to span selection for the features concept code, context, and status. The procedure that we used for calculating Kappa is described in detail by [9].

IAA for annotation of texts in the biomedical domain have been reported as Kappa as in [10, 11] and as F-measure described in [12] by [5, 13, 14]. [6] show that positive specific agreement, balanced F-measure, and Kappa are equivalent as P(E) approaches zero. The calculation of IAA for a set of annotations can be varied in many ways by defining different match criteria. We report on the following match criteria: spans are identical, spans overlap, spans overlap and concepts match, spans overlap and contexts match, spans overlap and status's match, and spans overlap, concepts, contexts, and status's match.

This selection of match criteria reflects our attitude towards exact span matching versus approximate span matching. We consider the agreement of concept code, context, and status assignments to be far more important than having exact span agreement. By relaxing the span matching criteria to require only an overlapping span match we get a better idea of how well the annotators agree with respect to the other attributes.

Results

The final gold-standard data set consists of 1556 annotations from the master consensus sets from the trial and experimental sets of notes. A total of 658 unique concept codes were used in the concept code assignments of the annotations. Because sixty of the notes were annotated under uncontrolled circumstances (i.e. the annotators communicated with us and each other) we report only results on the annotations for the remaining 100 notes. Table 2 provides a summary of the annotation sets that were created for these 100 notes including the number of annotations in each annotation set and the number of hours it took the annotators to create them. Because it took a total of 185 hours to create A1, A2, A3, A4, C1, and C2 we conclude that the annotator time to create a single pairwise consensus set for 100 documents is about 92 hours or roughly one document per hour. Table 3 provides the IAA results as positive specific agreement and Table 4 shows the IAA results as Kappa for only those annotations that are considered matches with respect to overlapping span for the attribute's concept code, context, and status. A1, A2, A3, and A4 are compared as average 2-way IAA in both tables.

An estimate on the upper bound of system performance

One would not expect an NLP system to agree with a human generated gold-standard data set better than the humans agree with each other. As such, the IAA numbers reported here represent upper-bounds on system performance as measured against the gold-standard. The most important datum that best represents the overall consistency of the gold-standard on the entire annotation task is the agreement between C1 and C2, 74.6%, for the match criteria that requires the spans to overlap and the concept code, context, and status values to match shown in Table 3. Because the final gold-standard data set is the result of an additional consensus step based on C1 and C2 we would expect that the consistency of GS to be better than C1 and C2. Thus, we conclude that 74.6% is a conservative upper bound for complete system performance.

Table 4 - IAA as Kappa

	A1,A2,A3,A4			C1,C2		
	K	P(A)	P(E)	K	P(A)	P(E)
concept	82.6	82.7	0.5	89.9	89.9	0.5
context	75.4	90.0	59.2	84.5	93.4	57.1
status	82.8	92.2	54.3	88.8	94.7	52.3
concept + context + status	71.0	71.1	0.3	82.1	82.1	0.3

Another datum of importance is the positive specific agreement, 81.7%, between C1 and C2 for the match criteria that requires the spans to overlap and the concept codes to match. A related datum appears in Table 4 that reports Kappa on concept code agreement for those annotations that match with respect to overlapping spans for C1 and C2, 89.9%. However, this represents an inflated measure of concept code agreement because this computation of Kappa removes annotations that disagree with respect to span. This effectively removes annotations that disagree with respect to concept code because the tasks of span selection and concept code assignment are very closely related per the annotation guidelines. It follows that measuring concept code agreement on only those annotations that agree with respect to exact span matching is even

higher (= 95.0% between C1 and C2). Therefore, we believe 81.7% to be a more reasonable estimate of the upper bound for system performance on the task of normalized named entity recognition.

Because the tasks of assigning context and status to annotations is much less related to the task of span selection, we consider the Kappa results in Table 4 to be a better indication of how well the annotator's agree than the corresponding percentage agreement results in Table 3. Thus, we consider 84.5% and 88.8% as Kappa agreement between C1 and C2 for context and status, respectively, to be the fairest measure of annotator agreement for those attributes. However, the absolute percentage agreements of 93.4% and 94.7% probably represent a more appropriate upper bound on system performance.

Table 3 - IAA as positive specific agreement

compared annotation sets	compared attributes					
	spans exact	spans overlap	spans overlap +			
			concept	context	status	concept + context + status
A1, A2, A3, A4	75.7%	87.9%	72.7%	79.0%	80.9%	62.5%
C1, C2	81.4%	90.9%	81.7%	84.8%	86.0%	74.6%
C1, MMTx	42.3%	47.0%	42.3%	n/a	n/a	n/a
C2, MMTx	38.2%	44.1%	37.3%	n/a	n/a	n/a
GS, MMTx	39.8%	45.7%	39.5%	n/a	n/a	n/a

The effect of creating a pair-wise consensus set

We hypothesized that having two individuals annotating independently and creating a consensus set would result in a much more consistent set of annotations. For every match criteria that we examined, there is a marked improvement in IAA between pair-wise IAA of individuals and the corresponding pair-wise consensus-level IAA. Table 3 shows that the average pair-wise IAA for the match criteria that requires spans to overlap and the concept code, context, and status rises from 62.5% for individual annotation sets to 74.6% for consensus annotation sets, a 32.3% disagreement reduction.

The effect of pre-annotating with MMTx

We hypothesized that providing the annotators with annotations from a third party system, MMTx, would be a good way to improve the speed and consistency of the annotation task without introducing bias that favors our system. Unfortunately, Table 2 shows that the annotators given the MMTx annotations, A1 and A3, annotated slower than the other two annotators, A2 and A4. Both A1 and A3 complained that the existing annotations slowed them down because spurious annotations and multiple mappings for the same span made them consider more concept codes than they would have otherwise. There was also no clear trend that the MMTx annotation improved pair-wise IAA between individuals. Also, MMTx did not have any measurable impact on the final assignments made in the master consensus set. If MMTx was providing "better" concept codes, then one would expect this to be reflected in the agreement between GS and MMTx. Table 3 shows the percentage agreement with respect to overlapping spans and concept code matching for C1 and MMTx to be 42.3% and 37.3% for C2 and MMTx. The agreement between GS and MMTx is about halfway between these two points at 39.5%. This is the case for every other IAA number reported for MMTx in Tables 3 and 4. Thus, we conclude that while the MMTx annotations seemed to influence the annotations in C1, there was no clear benefit to introducing this bias.

Acknowledgments

We acknowledge annotators Barbara Abbott, Debra Albrecht, Pauline Funk, and Donna Ihrke for their fine work, and Tanya Hoskin, Serguei Pakhomov, and James Buntrock for their input.

References

[1] Cohen KB, Fox L, Ogren PV, Hunter L. Empirical data on corpus design and usage in biomedical natural language processing. AMIA Annu Symp Proc 2005:156-60.

[2] Ananiadou S, McNaught J. Text mining for biology and biomedicine. Boston: Artech House; 2006.

[3] Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. J Biomed Inform 2003;36(6):414-32.

[4] Ogren PV. Knowtator: A Protégé plug-in for annotated corpus construction. Proceedings of the Human Language Technology Conference of the NAACL, 2006;Companion Volume:273-275.

[5] Mani I, Hu Z, Jang SB, Samuel K, Kraus M, Phillips J, et al. Protein Name Tagging Guidelines: Lessons Learned. Comparative and Functional Genomics 2005;6(Numbers 1-2):72-76.

[6] Hripcsak G, Rothschild A. Agreement, the F-Measure, and Reliability in Information Retrieval. J Am Med Inform Assoc 2005;12(3):296-298.

[7] Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 1960;20(1):37-46.
[8] Carletta J. Assessing agreement on classification tasks: The Kappa statistic. Computational Linguistics 1996;22(2):249-254.
[9] Poesio M, Vieira R. A corpus based investigation of definite description use. Computational Linguistics 1998;24(2):183-216.
[10] Pakhomov SV, Coden A, Chute CG. Developing a corpus of clinical notes manually annotated for part-of-speech. Int J Med Inform 2006;75(6):418-29.
[11] Tateisi Y, Tsujii J. Part-of-Speech Annotation of Biology Research Abstracts. Proceedings of the 4th International Conference on Language Resource and Evaluation 2004:1267-1270.
[12] Manning CD, Schütze H. Foundations of statistical natural language processing. Cambridge, Mass.: MIT Press; 2002.
[13] Gaizauskas R, Demetriou G, Artymiuk PJ, Willett P. Protein structures and information extraction from biological texts: the PASTA system. Bioinformatics 2003;19(1):135-43.
[14] Morgan A, Hirschman L, Yeh A, C. M. Gene Name Extraction Using Flybase Resources. Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine 2003:1-8.

Addresses for correspondence

philip@ogren.info