# Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition

**Philip V. Ogren M.S., Guergana K. Savova Ph.D., Christopher G. Chute M.D. Dr.P.H.**
Division of Biomedical Informatics, Mayo Clinic College of Medicine
Rochester, MN, USA
E-mail: philip@ogren.info, savova.guergana@mayo.edu, chute@mayo.edu

**Abstract**

We report on the construction of a gold-standard dataset consisting of annotated clinical notes suitable for evaluating our biomedical named entity recognition system. The dataset 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 (90.9%), concept code (81.7%), context (84.8%), and status (86.0%) agreement. Complete agreement for span, concept code, context, and status was 74.6%. We found that creating a consensus set based on annotations from two independently-created annotation sets can reduce inter-annotator disagreement by 32.3%. We found little benefit to pre-annotating the corpus with a third-party named entity recognizer.

## 1. Introduction

The text analysis group at the Mayo Clinic 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 (Pakhomov, Buntrock et al. 2005; Savova, Kipper-Schuler et al. 2008). By creating a gold-standard dataset 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 identifies. We describe the construction of this gold-standard corpus and evaluate its quality.

Cohen et al. (Cohen, Fox et al. 2005) 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 Ananiadou and McNaught (Ananiadou and McNaught 2006). 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 our current 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.

## 2. Materials and Methods

### 2.1. Annotation Task

#### 2.1.1. Clinical Notes Corpus

The Mayo Clinic has a repository of over twenty-five 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 randomly 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.

#### 2.1.2. 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. A span of text consists of two character offsets corresponding to the beginning and end of a selection of text. An annotation may have more than one span if the disorder mention cannot be reasonably captured by a single span. An annotation's concept code is a string attribute that contains a concept identifier from a controlled vocabulary, in this case SNOMED-CT.

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

[2] For this reason, our corpus will not be made publicly available in its current form. Contact the second author for further information.

The context of an annotation is a string attribute selected from the following list: *current*, *history of*, and *family history of*. The context attribute provides a way to capture whether the mentioned disorder is being considered in the present for the patient or if it is found in the context of the patient's personal or family medical history. The status of an annotation is a string attribute selected from the following list: *confirmed*, *possible*, and *negated*. The status value of an annotation is confirmed if the patient has the mentioned disorder, possible if it is undetermined whether the patient has the disorder, and negated if the patient does not have the disorder. 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. The following sentence serves to illustrate our annotation schema: "The patient returns with no complaints worrisome for recurrent metastatic oropharynx cancer." The disorder mentioned by "metastatic oropharynx cancer" maps to the concept *Stage IV Oropharyngeal Carcinoma* which has the concept unique identifier (CUI) of *C1378462*[3]. The disorder mention's context is *history of* and its status is *negated*.

An annotation may also be flagged as unrelated to the patient. Many times a mentioned disorder in a clinical note has little or nothing to do with the patient's health. For example, an unrelated disorder mention may appear in an organization's name (e.g. "Diabetes Clinic"), patient education (e.g. "patient was given a pamphlet on diabetes"), or medication side effects (e.g. "hyponatremia is quite common while taking this medication"). When an annotation is flagged as unrelated to the patient, the context and status are both automatically given the value *unrelated to patient*. The flag has no effect on the concept code assignment.

#### 2.1.3. Disorder concepts in SNOMED-CT

We created a subset of SNOMED-CT that contains only those concepts corresponding to disorders by leveraging the Unified Medical Language System (UMLS)[4] 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 (Bodenreider and McCray 2003)[5]. 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[6].

[3] For mapping to the concept unique identifier (CUI), consult http://kswebp2.nlm.nih.gov/UMLSKS/

[4] http://www.nlm.nih.gov/pubs/factsheets/umls.html. Version 2005AC

[5] We excluded *Finding* and *Signs and Symptoms* to constrain our task.

[6] We used the RRF Browser which is documented at http://www.nlm.nih.gov/research/umls/meta6.html

| 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 |

Table 1. UMLS Semantic Types used to subset SNOMED-CT. The type unique identifier (TUI) and the name of the type are given for each of the types we used.

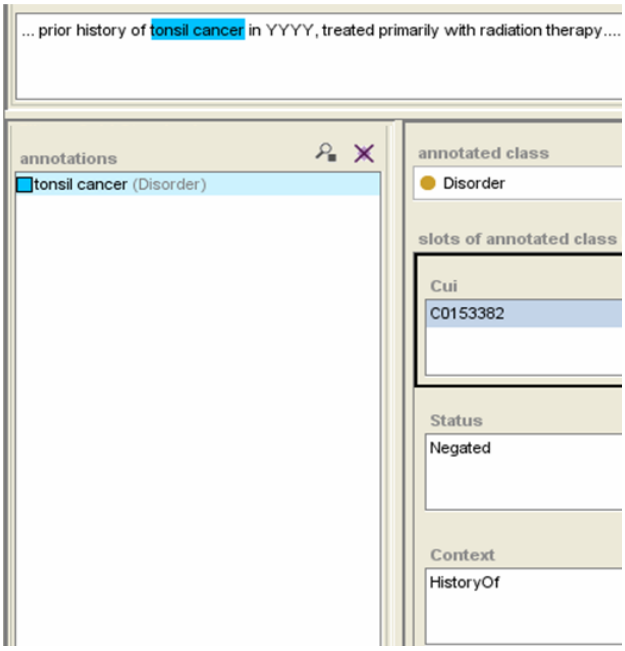


Figure 1: Knowtator screenshot with an example annotation

### 2.2. Gold Standard Corpus Construction

#### 2.2.1. Annotators

Four clinical data retrieval experts performed the annotation task. At the time of corpus construction, each of them had been in their current work positions for more than four years in addition to having prior experience in 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.

#### 2.2.2. Annotation Software

Knowtator[7] is a general purpose text annotation tool for creating gold-standard training and evaluation corpora for Natural Language Processing (NLP) systems described in

[7] http://knowtator.sourceforge.net/

(Ogren 2006). The annotation schema described in §2.1 is quite simple and was easily instantiated in Knowtator. 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 builds an initial consensus set based on two or more sets of annotations by consolidating completely matching annotations between the sets. All other annotations remain unchanged in the consensus set and must be manually reviewed and consolidated by the annotators. The end result is a single set of annotations that represents complete consensus agreement between the annotators. Figure 1 shows a screenshot of Knowtator as applied for this annotation task. The image shows a snippet of example text and an annotated disorder mention for the text "tonsil cancer" along with the attributes of CUI, status and context for the annotation.

**2.2.3. Pre-annotation of the corpus with MMTx**

We pre-annotated all 160 notes in the corpus using the MetaMap Transfer (MMTx) tool[8]described in (Aronson 2001). 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 §2.1.3. 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.

**2.2.4. Annotation Workflow**

The 160 notes were split into two sets: a trial set and an experimental set. The trial set consists of 60 notes and was used to help us better understand the annotation task. During the annotation of these 60 notes, the annotation schema was finalized, the annotation guidelines were completed, the annotators were trained, and frequent meetings were held to provide instruction to the annotators and answer their questions. Specific examples from the 60 notes being annotated were reviewed and the annotators were allowed to communicate with each other about the annotation task to a limited degree. The inter-annotator agreement results presented in this paper exclude annotations from this set of 60 notes.

Each of the 60 notes in the trial set 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 each of the annotators completed individual annotation of the 60 notes, a consensus set was created by A1 and A3 using the annotations they created individually by having the two annotators work together to resolve all differences. Similarly, a consensus set was created from A2 and A4's annotations by A2 and A4. By the time both pair-wise consensus sets were complete, the annotation schema and guidelines were finalized. At this point, a master consensus set consisting of 392 annotations was created based on both pair-wise consensus sets by all four annotators working together.

The experimental set consists of the remaining 100 notes and was also annotated individually by each of the four annotators. During this phase of the project 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. Again, A1 and A3 were given MMTx annotations and A2 and A4 were not. Similarly, pair-wise consensus sets were created for A1 and A3's annotations and for A2 and A4's annotations, respectively. Finally, a master consensus set consisting of 1164 annotations was created from both pair-wise consensus sets by all four annotators working together.

**2.2.5. Annotation Guidelines**

(Mani, Hu et al. 2005) 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 (paraphrased):

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) Annotate all mentions of disorders in each note. 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.

[8] http://mmtx.nlm.nih.gov/

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 (see §2.1.2), 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?*

### 2.3. Inter-Annotator Agreement

To calculate IAA we used positive specific agreement described by (Hripcsak and Rothschild 2005). We used the following match criteria for calculating IAA:

- spans are identical
- spans overlap
- spans overlap and concept codes match
- spans overlap and contexts match
- spans overlap and statuses match
- spans overlap and concept codes, contexts, and statuses match.

Calculating Kappa can be problematic because expected chance agreements when span comparisons are involved are very small. However, we did calculate Kappa statistic described in (Cohen 1960) and (Carletta 1996) for the attributes concept, context, and status for annotations where spans already agree using a procedure similar to the one described in (Poesio and Vieira 1998).

## 3. 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 the notes in the trial set were annotated under uncontrolled circumstances (i.e. the annotators communicated with us and each other) we report only results on the annotations for the 100 notes in the experimental set. Table 2 provides a summary of the annotation sets that were created for the experimental set of notes including the number of annotations in each annotation set and the number of hours it took the annotators to create them. It took a total of 184.5 hours to create A1, A2, A3, A4, C1, and C2. Therefore, we conclude that the annotator time to create a single pair-wise consensus set for 100 documents is about 92 hours or roughly one document per hour. Table 3 provides the IAA results as percentage agreement and Table 4 shows the IAA results in terms of Kappa for only those annotations that match with respect to overlapping span for the attribute's concept code, context, and status. For example, Table 3 shows that 47.0%, or 840 annotations of the 1789 annotations in annotation sets C1 and MMTx (see Table 2), are matches with respect to the overlapping span criteria. The Kappa measurements for agreement between C1 and MMTx shown in Table 4 are calculated on only a subset of the 840 annotations that match with respect to overlapping spans.

### 3.1. 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 C3 to be better than C1 and C2. Thus, we conclude that 74.6% is a conservative upper bound for complete system performance, i.e. we expect that the true upper bound is likely higher.

Another datum of importance is the percentage 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 removing annotations that disagree with respect to span effectively removes annotations that disagree with respect to concept code because the tasks of span selection and concept code assignment are very closely related. It is no surprise, then, that measuring concept code agreement on only those annotations that agree with respect to *exact* span matching is even higher ($\kappa$ = 95.0% for 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.

| annotation set | description | Count (number of annotations) | Hours |
|---|---|---|---|
| A1 | pre-annotated with MMTx | 1105 | 53.5 |
| A2 | not pre-annotated | 1142 | 53 |
| A3 | pre-annotated with MMTx | 1054 | 38 |
| A4 | not pre-annotated | 1113 | 21 |
| C1 | consensus of A1 and A3 | 1125 | 8 |
| C2 | consensus of A2 and A4 | 1193 | 11 |
| C3 | consensus of C1 and C2 | 1164 | 13 |
| MMTx | annotations created by MMTx | 664 | <1 |

Table 2: The number of annotations for each annotation set created for the experimental set of notes is shown with the number of hours it took to create the annotation set.

| | **Compared attributes** | | | | | |
|---|---|---|---|---|---|---|
| | | | spans overlap + | | | |
| **Compared annotation sets** | spans exact | 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 |
| C3, MMTx | 39.8% | 45.7% | 39.5% | n/a | n/a | n/a |

Table 3: IAA as percentage agreement on all annotations in compared sets. The overall agreement between C1 and C2 when the match criteria requires that the spans overlap and the concept, context, and status are the same is 74.6%. Comparison between A1, A2, A3, A4 is an average of 2-way agreement.

| | | compared attributes | | | |
|---|---|---|---|---|---|
| compared annotation sets | | concept code | context | status | concept code + context + status |
| A1, A2, A3, A4 (avg. 2-way agreement) | κ | 82.6% | 75.4% | 82.8% | 71.0% |
| | P(A) | 82.7% | 90.0% | 92.2% | 71.1% |
| | P(E) | 0.5% | 59.2% | 54.3% | 0.3% |
| C1, C2 | **κ** | **89.9%** | **84.5%** | **88.8%** | **82.1%** |
| | P(A) | 89.9% | 93.4% | 94.7% | 82.1% |
| | P(E) | 0.5% | 57.1% | 52.3% | 0.3% |
| C1, MMTx | κ | 89.9% | n/a | n/a | n/a |
| | P(A) | 90.0% | | | |
| | P(E) | 0.7% | | | |
| C2, MMTx | κ | 84.5% | n/a | n/a | n/a |
| | P(A) | 84.6% | | | |
| | P(E) | 0.7% | | | |
| C3, MMTx | κ | 86.3% | n/a | n/a | n/a |
| | P(A) | 86.4% | | | |
| | P(E) | 0.7% | | | |

Table 4: Kappa is calculated on only those annotations that match with respect to overlapping spans for the respective compared annotation sets.

### 3.2. The effect of creating a pair-wise consensus set

Because we were uncertain that an individual annotator could make a high quality set of annotations by herself on a single pass, 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. For example, 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 34% disagreement reduction. It is also interesting to note that even for least strict match criteria that requires only that the spans overlap where initial agreement is highest 87.9% the agreement of the consensus sets for the same criteria is much greater at 90.9%. This represents a 25.4% disagreement reduction. That is, individual annotators are quite good at "disorder spotting" on their own but still benefit substantially from having a second annotator annotate the document.

### 3.3. 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 a 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. The IAA results between A1 and A3 were, in general, higher than those between A2 and A4. However, this trend is confounded by the fact that the consistently highest agreement was between A1 and A2 and that A4 had consistently lower IAA results with each of the other annotations sets, including C1, C2, and C3. These confounding trends can reasonably be explained by the lengths of time spent on the annotation task by the respective annotators and by comparing the educational backgrounds of the annotators (e.g. A1 and A2 have the most similar work experience.)

Finally, MMTx did not have any measurable impact on the final assignments made in the master consensus set. If the MMTx was providing better concept codes, then one would expect this to be reflected in the agreement between C3 and MMTx. That is, the agreement between C3 and MMTx should be closer to that of C1 and MMTx than C2 and MMTx. The data does not bear this out. For example, 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 C3 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.

## 4. Discussion

One of the major concerns voiced by the annotators was the difficulty of navigating the large subset of SNOMED-CT that was given them. The sheer number of concepts (over 82,000) made for a daunting search space. We hypothesize that it will be possible to characterize the SNOMED-CT codes that were easy to agree on versus ones that were not similar to the analysis on Gene Ontology annotations in (Ogren 2005). Such analysis will hopefully point us towards ways to improve annotation consistency. It may also prove useful to relax the concept code matching requirement in a way that exploits the hierarchical relationships of SNOMED-CT. It may be that most of the concept code assignment disagreements are very small with respect to some distance metric defined by the relationships in SNOMED-CT so as to be irrelevant.

We observed that a main source of disagreement was in the attribute assignments. One very frequent set is the conditional sentence construction like "If the patient develops tachycardia…", which led the annotators to assign different status values to the disorder mention annotation for "tachycardia". The status *possible* was given with the reasoning being that it is possible to develop the disorder. The status *negated* was given with the reasoning being that if the patient could develop a disorder, then she must not have it. Another frequent set of disagreements stemmed from the context value assignments. For example, in "history of diabetes", one annotator assigned the context *current* to diabetes disorder mention with the reasoning being that a diabetes diagnosis is for life, hence the context value cannot be *history of*.

Another set of disagreements is due to span decisions. For example, in "black, tarry, bloody stools", the concept identifier C025222 can be assigned to either "black stools", "tarry stools" or "black, tarry stools" which led to span variations.

We also noticed that some semantic types which are included in the Disorder definition (see Table 1), e.g. Pathological Function and Injury or Poisoning, map to terms that one could argue are not strictly disorders. The following examples give the text of a disorder mention followed by the term an annotator mapped it to along with its semantic type of the term:

- healing problems → impaired wound bleeding (Pathologic Function)
- obstructing → obstruction (Pathologic Function)
- smoking →tobacco dependence (Mental or Behavioral Dysfunction)
- side effects → adverse effects (Pathologic Function)
- domestic violence → domestic violence (Mental or Behavioral Dysfunction)

Although these terms belong to the semantic types that we used for defining disorders, they are at best ambiguous as to whether they are, in fact, disorders. In some cases the annotators adhered to the strict definition of a disorder as any concept that belonged to the set of concepts that we provided them while in other circumstances the annotators used their best judgment to filter out unlikely candidates.

One semantic type that was excluded from the list of semantic types for disorders given by (Bodenreider and McCray 2003) was *Signs and Symptoms*. However, the distinction between concepts that fall in this semantic type and concepts that were included in the annotation task can often be a fine one. Some examples of terms that correspond to semantic types other than *Signs and Symptoms* but seem to belong to that semantic type are: "blood in stool", "inflamed tonsils", "anxious", and "congested." Terms such as these were often ignored by one annotator and mapped to a concept by another.

A final source of disagreements that we discuss here relate to structural properties of the documents. Clinical notes often contain headings for exams performed on a specific body part. Some of the annotators chose to include the heading as part of the named entity while others did not. For example, for the sentence "Lymph: No palpable adenopathy in the cervical, supraclavicular, axillary, or inguinal node chains", one annotator marked the disjoint span "Lymph….adenopathy" and thus included the heading as part of the named entity. Another annotator chose only the textual mention "adenopathy" as the disease named entity. While our guidelines allowed for discontinuous spans it was not clear which annotation most faithfully adhered to the rules.

While it may seem that this data provides an implicit evaluation of the MMTx system, it is important to note that this is true in only a very limited sense. A fair evaluation for MMTx would attempt to maximize the F-measure by adjusting the relevance score that MMTx provides for each mapping (see §2.2.3). No such experimentation was conducted. For our use of MMTx, recall was sacrificed at the expense of improved precision.

## 5. Conclusions

We have described the construction of a gold standard evaluation corpus for evaluating our clinical named entity recognition system and quantified its quality using IAA metrics. We found that pair-wise annotation with a subsequent round of consensus annotation results in a large reduction in disagreement but that pre-annotating the text with a third party system was not helpful. A companion LREC 2008 manuscript uses this corpus for system evaluation of the Mayo Clinic Named Entity Recognition system (Kipper-Schuler, Kaggal et al. 2008).

## 6. Acknowledgements

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

## 7. References

Ananiadou, S. and J. McNaught (2006). "Text Mining for Biology and Biomedicine." Computational Linguistics **33**(1).

Aronson, A. R. (2001). "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." Proc AMIA Symp **17**(21): 17-21.

Bodenreider, O. and A. T. McCray (2003). "Exploring semantic groups through visual approaches." Journal of Biomedical Informatics **36**(6): 414-432.

Carletta, J. (1996). "Assessing agreement on classification tasks: the kappa statistic." Computational Linguistics **22**(2): 249-254.

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales." Educational and Psychological Measurement **20**(1): 37.

Cohen, K. B., L. Fox, et al. (2005). "Empirical data on corpus design and usage in biomedical natural language processing." AMIA Annu Symp Proc: 156-60.

Hripcsak, G. and A. S. Rothschild (2005). "Agreement, the F-Measure, and Reliability in Information Retrieval." Journal of the American Medical Informatics Association **12**(3): 296-298.

Kipper-Schuler, K., V. Kaggal, et al. (2008). "System evaluation on a named entity corpus from clinical notes." Proc. LREC.

Mani, I., Z. Hu, et al. (2005). "Protein name tagging guidelines: lessons learned." Comparative and Functional Genomics **6**(1-2): 72-76.

Ogren, P. V. (2005). "Implications of Compositionality in the Gene Ontology for its Curation and Usage." Biocomputing 2005: Proceedings of the Pacific Symposium, Hawaii, USA 4-8 January 2005.

Ogren, P. V. (2006). "Knowtator: A Protégé plug-in for annotated corpus construction." Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations: 273-275.

Pakhomov, S., J. Buntrock, et al. (2005). "High Throughput Modularized NLP System for Clinical Text (Interactive Poster)." 43rd Annual

Meeting of the Association for Computational Linguistics, Ann Arbor, MI.

Poesio, M. and R. Vieira (1998). "A Corpus-based Investigation of Definite Description Use." Computational Linguistics **24**(2): 183-216.

Savova, G., K. Kipper-Schuler, et al. (2008). "UIMA-based Clinical Information Extraction System." Proc. UIMA for NLP Workshop. LREC