System evaluation on a named entity corpus from clinical notes

Karin Kipper-Schuler, Vinod Kaggal, James Masanz, Philip Ogren and Guergana Savova
Division of Biomedical Informatics, Mayo Clinic College of Medicine
Rochester, Minnesota, USA
E-mail: {schuler.karin, kaggal.vinod, masanz.james, savova.guergana}@mayo.edu

Abstract

This paper presents the evaluation of the dictionary look-up component of Mayo Clinic's Information Extraction system. The component was tested on a corpus of 160 free-text clinical notes which were manually annotated with the named entity disease. This kind of clinical text presents many language challenges such as fragmented sentences and heavy use of abbreviations and acronyms. The dictionary used for this evaluation was a subset of SNOMED-CT with semantic types corresponding to diseases/disorders without any augmentation. The algorithm achieves an F-score of 0.56 for exact matches and F-scores of 0.76 and 0.62 for right and left-partial matches respectively. Machine learning techniques are currently under investigation to improve this task.

1. Introduction

The natural language processing (NLP) field has a vast repository of well annotated data which allowed both rule-based and statistical systems to be developed and compared. The biomedical field also has many such resources available. [1], [2] These biomedical repositories mostly contain literature abstracts and genomic data. Resources such as these are invaluable for evaluation and improvement of software systems.

Clinical data describing encounters between physicians and patients, on the other hand, are a rare commodity. In order to protect patient privacy, these corpora are generally only available locally in medical institutions and cannot be widely distributed. The only publicly available corpus to our knowledge is the one released by Cincinnati's Children's Hospital Natural Language Processing group (Pestian et al., 2007). This corpus is annotated at the document level for billing codes (ICD-9[3].)

At Mayo Clinic we have an NLP-based Information Extraction system which annotates patient records for named entities (NE) such as anatomical sites, drugs, diseases/disorders, signs/symptoms, procedures, and level of activity. This annotated data is used internally to support research such as the identification of patients who are good candidates for clinical trial studies. In order to evaluate the named entity recognition (NER) component of our system, specifically the disease/disorder entities, we manually annotated a corpus consisting of 160 free-text clinical reports which are mostly transcriptions of physician-patient encounters.

This paper addresses the evaluation of the NER component in the Mayo Clinic Information Extraction system. In Section 2 we describe the corpus and point out some of the language challenges associated with this type of documents. In Section 3 and Section 4, we present the algorithm and the metrics used in the evaluation. In Section 5 we discuss the results and the limitations of our approach. Finally, in Section 6 we describe other work currently being done within our system and outline future directions.

2. Corpus description

The corpus used for the evaluation is comprised of 160 randomly selected clinical notes. Clinical text has its own characteristics which set it apart from other types of text such as newswire (Pakhomov et al., 2006.) Clinical notes are textual descriptions of physician-patient encounters and frequently include incomplete sentences, inverted constructions, misspellings and spelling variations, and a heavy use of abbreviations and acronyms.

These notes were annotated for disease/disorder NEs as defined by a collection of concepts belonging to the Unified Medical Language System ontology (UMLS.) [4] From that ontology, we focused on the semantic types listed in Table 1 (Bodenreider and McCray, 2003).

Semantic Type UMLS ID	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 corresponding to

[1] http://bioie.ldc.upenn.edu/wiki/index.php/Main_Page
[2] http://compbio.uchsc.edu/ccp/corpora/obtaining.shtml
[3] http://www.cdc.gov/nchs/icd9.htm
[4] http://www.nlm.nih.gov/pubs/factsheets/umls.html

diseases/disorders

These 160 clinical notes were manually annotated independently by four medical retrieval specialists who together reviewed disagreements to create the annotations gold standard (Ogren et al., 2008.) Inter-annotator agreement was measured as Kappa (Carletta, 1996; Poesio and Vieira, 1998) with the range of 0.71-0.899.

Each disorder/disease NE identified was annotated with attributes:

- The UMLS Concept unique identifier (CUI) of the disease
- Status (confirmed, possible, negated)
- Context (current, history_of, family_history_of)
- Unrelated to patient (true, false)

The guidelines for the manual annotations permitted the experts to mark disease entities not only according to what is explicitly present in the ontology, but to use their best judgment in capturing reasonable synonyms of the concept within the ontology. All mentions of diseases were annotated in each report, independent of whether they belonged to the patient. Only a single annotation per mention of disease/disorder was allowed (nested annotations such as *heart disease* and *disease* were not allowed.) The annotation which corresponded to the most specific concept was preferred, and disjoint spans of annotations, such as "Edema … legs", were allowed (to capture non-contiguous spans of text which refer to a more specific concept.)

Some examples pointing to the difficulty in selecting synonyms are shown in the following pairs of manual annotations and their mappings to ontology concepts: "side-effects from the medication – adverse drug effect", "elevated blood pressure -- hypertensive disease", and "illness – physical illness". Recognizing "illness" as physical illness is difficult because it needs to be differentiated from terminal illness, which is also a disorder concept in UMLS.

3. Dictionary look-up algorithm in the Mayo Clinic Information Extraction system

Each report was processed through the Mayo Clinic Information Extraction system which is being used to process and extract information from free-text clinical notes. Its main function is the discovery of clinical named entities such as diseases, signs/symptoms, medications, anatomical sites and procedures. Attributes related to named entities – context, status and relatedness to patient – are also extracted from the text.

The system consists of several annotators such as context-free tokenizer, context-sensitive spell corrector, lexical normalizer annotator, sentence detector, context dependent tokenizer, part-of-speech tagger, shallow parser, named entity recognizer, and negation detector based on NegEx (Chapman et al., 2001.) The system is built upon the Unstructured Information Management[5] (UIMA) framework. For details on the system see (Pakhomov et al., 2005; Savova et al. 2008.)

Unlike the manual annotation, the system outputs all named entities recognized. These may include multiple annotations for the same span:

There is no history of [peptic ulcer disease].
There is no history of [peptic ulcer] disease.
There is no history of peptic [ulcer disease].
There is no history of peptic ulcer [disease].

The current study evaluates a dictionary look-up algorithm. The dictionary – in our case, a subset of UMLS as described above – was expanded with synonyms and indexed by the heads of the noun phrases. Based on the output of a shallow parser, the algorithm finds all noun phrases and their respective heads within a clinical note. Permutations of variations of the head and modifiers are looked for in the dictionary. Our goal is to evaluate this algorithm, understand the main sources of errors and suggest improvements.

4. Evaluation metrics

We used several metrics to evaluate the output from the NER module according to (Tsai et al., 2006.)

The exact match criteria checks whether the spans of the manually annotated disease and the spans from the NER module are exactly the same. We computed recall, precision and F-score as our metrics:

	Gold Standard		
	True	False	
Positive	True positive	False positive (Type I error)	Positive Predictive Value
Negative	False negative (Type II error)	True negative	Negative Predictive Value

Table 2: Evaluation metrics

(1)

$$recall = \frac{numberOfTruePositives}{numberOfTruePositives + numberOfFalseNegatives}$$

(2)

$$precision = \frac{numberOfTruePositives}{numberOfTruePositives + numberOfFalsePositives}$$

(3) $$F\_score = \frac{2 * Precision * Recall}{Precision + Recall}$$

[5] http://incubator.apache.org/uima/

For exact matches, we computed two types of results: 1) using only the spans, and 2) additionally taking into consideration the attributes of each annotation.

Because the NER module allows multiple annotations per span, nested spans, and overlapping spans, we also computed precision, recall and F-score for the following types of partial matches. This measure has been referred to as soft F1-score in (Settles, 2005.)

- Right boundary matches: the end offset of the NE found by the system matches the end offset of the manual annotation (subsumed and not subsumed by exact match), e.g., gold standard annotation is "metastatic cancer", and NER output is "cancer".
- Left boundary matches: the beginning offset of the NE found by the system matches the beginning offset of the manual annotation (subsumed and not subsumed by exact matches), e.g., gold standard annotation is "carcinoma of the liver", and NER output is "carcinoma".
- Partial matches (inside matches subsumed or not subsumed by exact matches), e.g., gold standard annotation is "metastatic cancer of the liver", and NER output is "cancer".

We evaluated the accuracy of negation detection separately.

5. Results and discussion

Figure 1 shows the number of annotations in each category. The total number of manual annotations is 1,957 and the system identifies 1,820 annotations. The gold-standard and the system have a total of 1,601 annotation in common (this number includes exact and partial matches).

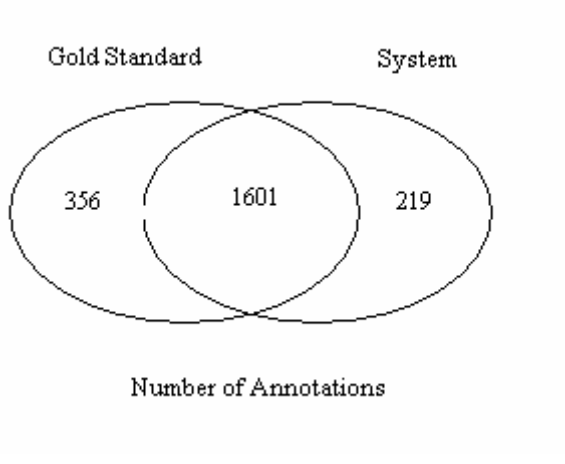

Figure 1: Venn diagram for gold standard and system annotations

Evaluation results are summarized in Table 2 and Table 3. The results from exact matches improve when boundary and non-boundary matches are taken into consideration. The best result achieves an F-score of 0.81.

Negation detection performs remarkably well (accuracy = 0.94). The concept "cancer" in the next example receives a negation mark. "There is no visible or palpable mucosa abnormality worrisome for cancer." Assigning the correct negation value in this case requires a degree of inferencing which the algorithm currently lacks.

As expected, CUI accuracy is high for exact span matches. It drops when all possible matches are considered. For the gold standard annotation of "degenerative joint disease" with CUI C0029408, the algorithm produces the right boundary match "joint disease" with CUI C0022408. As the boundary match is a broader term than the manual annotation, its CUI is that of the parent of "degenerative joint disease". In our evaluation, this is considered as a CUI non-match.

	Recall	Precision	F-score
Exact matches (k=0.81)	0.63	0.51	0.56
Exact matches and Right boundary matches (k=0.90)	0.74	0.80	0.76
Exact matches and Left boundary matches (k=0.90)	0.67	0.56	0.62
Exact matches and Partial non-boundary matches (k=0.90)	0.65	0.53	0.59
All Matches (k=0.90)	0.76	0.88	0.81

Table 2: Evaluation results – recall, precision, F-score. Results in brackets are relevant kappa values.

	Accuracy negation	Accuracy CUI
Exact matches		0.95 (k=0.90)
All Matches	0.94 (k=0.89)	0.56 (k=0.90)

Table 3: Evaluation results – accuracy. Results in brackets are relevant kappa values.

5.1 Sources of Errors

We can categorize the sources of errors broadly as textual errors, algorithmic errors, manual annotation problems, dictionary problems, and finally conceptual problems.

Spelling mistakes such as "bulemia" found in the text are most likely due to the transcription. Other sources of textual errors include incorrect assignment of part-of-speech tags and/or incorrect chunk/parse which causes the look-up to fail as the lookup algorithm window

is restricted to noun phrases.

Many of the algorithm errors are due to abbreviations and ambiguity. We had to filter out abbreviations such as "Dr." and "MD" which could potentially be used as abbreviations for "diabetic retinopathy" and "Duane retraction syndrome", and "dysmyelopoietic syndromes", "Miller Dieker syndrome", respectively. Another source of algorithm errors is due to the annotation of disjoint spans, i.e., non-contiguous words. Examples of such annotations include "Lymph: No adenopathy in the neck or axilla", with "Lymph…adenopathy" marked with the concept "Disorder of lymph node." (and with the negated attribute). Disjoint span annotations accounted for a large number of the manually annotated named entities (186 cases.) These were included by the experts in order to convey the most specific concept possible. At present the algorithm does not handle cases of disjoint annotations.

Another source of errors is missed named entity annotations in the gold standard. Some examples discovered by the algorithm are "hematuria", "tetanus", "fever" and "heaves".

Lexical variation is another source of errors e.g. the text "bladder showed very mild trabeculation" was mapped to the disorder "trabeculated bladder". In order to retain the concepts, extremely difficult concepts were included in the manual annotations. Such cases include "Spinal stenosis L3 through L5" and "L3, 4, 5 interspaces ... stenosis" both mapping to the concept "Spinal stenosis of lumbar region", and "left hilar ... right hilar adenopathy" mapping to "Bilateral hilar adenopathy".

The task of keeping up with a wide-coverage resource as a knowledge base for the lookup algorithm is very expensive and time-consuming. Our dictionary was derived automatically from SNOMED. It includes entries with conflated disorders such as "Myalgia or myositis NOS (disorder)" and problematic entries such as "Fungal infectious disease, NOS." We are in the process of cleaning up the dictionary entries and augmenting the original dictionary with variants. However the examples from the previous paragraph beg the question of how many variants we would need to add to our dictionary to be able to handle such cases using the current approach.

Certain semantic types shown in Table 1 were found to be too general to capture disorder/disease meaningfully. In particular, semantic type *T046 pathologic function* seems to be a catch all category – "allergies", "congested", "complication", "side effects", and "free fluid" all map to that category; and semantic type *T020 acquired abnormality* includes procedures such as "abdominal hysterectomy." We also question the utility of discovering very broad disorder NEs, e.g. "disorder", "illness", "mass", "nodules". Such NEs could be left out from the dictionary by including only terms that are somewhat removed from the ontology root.

5.2 Extending the dictionary look-up approach

Although a dictionary look-up approach has the potential to work well, the characteristics of clinical text which includes many linguistic variations, constant introduction of new terms, disjoint concepts, and extensive use of abbreviations and acronyms adds to the complexity of the task.

We are in the process of investigating the use of machine learning techniques for NER applied to our domain. Currently, we are performing experiments with Conditional Random Fields (CRFs) (Lafferty et al., 2001; McCallum and Li, 2003) and Support Vector Machines (SVMs) (Ji et al., 2002.) We experimented with several features as both techniques support multiple features during learning. CRFs' main strength lies in their ability to include various unrelated features, while SVMs' advantage is in the inclusion of overlapping features.

Our results show that CRFs with multiple features significantly outperforms a single feature of dictionary look-up which we used as a baseline for that evaluation. The highest performance, with an F-score of 0.86, is achieved when orientation, context window size of 3, and capitalization are used as features in addition to the dictionary look-up.

Puzzling, however, is the fact that in our experiments with SVMs the combination of features that performed well with CRFs did not provide any increase in performance to the dictionary look-up baseline.

6. Conclusion

Our results show the complexity of identifying disease/disorder mentions in a corpus of clinical reports. Because the dataset was annotated by experts with vast knowledge of the domain, many of the manual annotations go beyond simple dictionary entries and their morphological and syntactic variants. The manual annotations allowed a complex level of synonymy which we believe relevant and that we would like to capture to enrich our medical research.

Pursuing the path of improving the automatic annotations and keeping up with the diversity of language in clinical text; we are currently investigating the use of machine learning techniques to identify disease mentions in our corpus. We have been experimenting with Conditional Random Fields and Support Vector Machines as applied to named entity recognition.

We are also planning to augment this corpus by annotating other clinically-relevant named entities such as drugs information and side-effects, and smoking status, and perform evaluation in the same fashion.

Acknowledgements

We would like to thank the meticulous work of our retrieval experts Barbara Abbot, Debra Albrecht, Pauline Funke, and Donna Ihrke in annotating the corpus.

7. References

Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. Journal Biomedical Informatics 2003; 36(6):414-32

Carletta J. Assessing agreement on classification tasks: The Kappa statistic. Computational Linguistics 1996; 22(2):249-254.

Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal Biomedical Informatics. 2001; 34:301-10.

Ji K, Maino T, Ohta Y, Tsujii Ji. Tunning Support Vector machines for biomedical named entity recognition. NLP in the Biomedical Domain Workshop. ACL; Philadelphia; 2002.

Lafferty A, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001), 2001.

McCallum A, Li W. Early Results for Name Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. ACM Transactions on Computational Logic, Vol.V, No. N; 2003.

Ogren P, Savova GK and Chute, CG. Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition. LREC 2008, Morocco.

Pakhomov, S., Buntrock, J., Duffy, P.. High Throughput Modularized NLP System for Clinical Text (Interactive Poster). In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005), MI; 2005.

Pakhomov S, Coden A. and Chute CG. (2006). Developing a corpus of clinical notes manually annotated for part-of-speech. International Journal of Medical Informatics (2006) 75, 418-429.

Pestian JP, Brew C, Matykiewicz PM, Hovermale DJ, Johnson N, Cohen KB, Duch W. A shared task involving multi-label classification of clinical free text. Proceedings of ACL BioNLP; Prague; 2007.

Poesio, M. and Vieira, R. A corpus-based investigation of definite description use. Computational Linguistics1998; 24(2), 183-216.

Savova GK, Kipper-Schuler KC, Buntrock JD, Chute CG. UIMA-based Clinical Information Extraction System. UIMA Workshop, LREC, Morocco; 2008.

Settles B. ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text. Bioinformatics, 21(14):3191-3192., 2005.

Tsai RT, Wu SH, Chou WC, Lin YC, He D, Hsiang J, Sung TY, Hsu WL. Various Criteria in the Evaluation of Biomedical Named Entity Recognition. BMC Bioinformatics, 7:92. 2006.