Natural Language Processing Framework to Abstract Problems, Treatments, and Tests from Clinical Documents

Henry Ware[1], Charles J. Mullett, MD, PhD[2], V. Jagannathan, PhD[1], Oussama El-Rawas[1]

[1]MedQuist, Morgantown, WV; [2]West Virginia University, Morgantown, WV

Abstract

We present an NLP framework to abstract problems, treatments and tests from clinical documents. The framework was developed in the context of the Fourth i2b2/VA NLP Challenge. The challenge this time was broken into three levels. The first level focused on concept recognition and the other two levels looked at the truth value of the concepts and relationship among them. MedQuist focused on the first part of the challenge which involved recognizing concepts representing problems, treatments and tests.

We developed our framework using a combination of open source modules optimized through rule sets developed to best fit the annotated training data available for 349 clinical documents. The Challenge's test set comprised of 477 clinical documents released to development teams on the day of competition. Our framework achieved an F-measure of 77.3% for exact class span of 77.3%, and 89.4% for inexact class span on the training set of annotated clinical documents. The F-measure for the same two categories for the 477 test documents released in the competition were: 73.4% & 86% respectively.

The framework and approach used to detect concepts is being incorporated in practical applications which includes as one of the steps, the human review of annotated concepts. These efforts show much promise to improve structuring of real-world clinical documentation.

Introduction

The HITECH act passed in February 2009 as part of the American Reinvestment and Recovery Act (ARRA) sets aside significant funds as incentives for the adoption of EMR. In particular, the act calls for the demonstration of "meaningful use of Certified Electronic Medical Record (EMR)" to qualify for financial incentives. Evidence of meaningful use has been defined, in part, as the capturing of structured elements in an EMR such as problem lists, medications, procedures, allergies, and quality measures.

The NLP challenge organized by i2b2 this year, with its focus on concepts that relate to problems, treatments and tests, is highly relevant to the general challenge of structuring such elements in EMRs.

The framework discussed in this report attempts to extract problems, treatments and tests found in discharge summaries and progress notes.

Background

There has been significant interest in Natural language processing (NLP) technology recently, as evidenced by the robust participation in the NLP-challenges over the past few years[1, 2, 3,4,,5]. There has also been a few evaluation of NLP from commercial vendors in terms of their ability to abstract medications[6,7]. NLP is a very active area of research in the healthcare community and the challenge organized by i2b2 and its other members have taken significant steps to address the paucity of annotated gold standards for various concept categories. By addressing one of the core enablers of improving research in this area, the organizers are doing a great service for the whole discipline. The effort described here outlines our approach to leverage this annotated content. In addition, we discuss our approach to incorporate various vocabulary and tools suites to craft together a framework that provides a comprehensive abstraction solution.

Methods

The process of building an NLP solution to abstract concepts has various components.. One of the first steps is to pull together a lexicon that catalogs the concepts of interest. Another step called by the challenge is the linguistic parsing of the text, as concept annotation called for the complete annotation of the phrase in which the concept appeared. Then we have the various approaches to recognizing a concept as a true concept in the context in which they appear. We focused on the rule based approach described here.

Lexicon preparation

This consumed significant time and involved cataloging a range of different concept categories. We used the Apelon solution to access publicly available vocabularies[8]: SNOMED, ICD9, LOINC. Lexicon preparation focused on the following categories:

- Signs and symptoms – culled from SNOMED. Example terms: abdominal pain, hip sores, necrotic ulcers, wound infection
- Diagnosis and findings – culled from SNOMED and ICD9
- Body parts / anatomical sites – culled from SNOMED
- Color catalog – culled from Internet search/Google and clinical documents
- Treatments – procedures – culled from SNOMED
- Treatments – medications – culled from RxNorm and Multum
- Tests – culled from LOINC, SNOMED and a medical transcriptionists resource from Interfix

Building each lexicon took some significant amount of processing (both automated and manual) as none of the vocabularies fully expressed terms in the manner that they actually appear in clinical documents.

Additional lexicons were collected for each of the categories from the training documents by a semiautomated process. These included abbreviations and variant spellings in addition to missing concepts.

Some of the training documents were created through a form-filling process at the original clinical institution. The labels from the forms themselves were mingled with the patient specific medical information. As these labels were in a rather limited set, they were collected into a special lexicon.

Document Processing

We used a rule based approach to the problem. We have had success with this approach in the past on more focused queries and were interested in how well it would scale a problem with this many distinct concepts. Time constraints obviously would not allow any marginal engineering time at all on the typical concept, so this approach is somewhat outside its sweet spot.

As one of the rules of the contest was to annotate full phrase structure, we adopted the cTakes solution from Mayo[11] built on top of the UIMA framework. The first step was to run all the documents through cTakes parts-of-speech tagging solution and use that as the basis for further processing.

Documents were tagged with the originating hospital, work type and source (dictation or other). This information was not available from metadata, so rules were used based on the document content. Documents were then divided into sections by regular expressions. The sections were categorized into groups including: headers, footers, medications, and labs.

Words matching any lexicon were identified by regular expression. These were expanded to phrases by a rule based system which considered part of speech.

Lexicon matches are not sufficient to categorize a term. Some terms are ambiguous. For example, the term "Potassium" is very ambiguous, appearing in all three senses in the training data. Potassium is a test in "Potassium is high", a problem in "high Potassium" and a treatment in "Potassium chloride". Sometimes a term had two different meanings in the same sentence: "Pt ambulated with PT in the halls yesterday". This sentence is clear--- the patient walked with the physical therapy team--- but not straight-forward, as "Pt" refers to "patient," and "PT" to "Physical Therapy." Likewise a "flat plate" is an implant, but "her plate" is more likely to be what she ate on.

Comment [Chuck Mul1]: Recheck this example. "Flat plate" is frequently used to describe an abdominal x-ray taken in the horizontal position. http://medical-dictionary.thefreedictionary.com/flat+plate

Even a term like "Congestive heart failure" which falls squarely in the problem category can be part of a treatment in "Congestive heart failure protocol". We call terms like "protocol" toggles, because they switch the category of a concept .

To categorize a term, a rule based procedure was followed. The matching lexicon(s) give a default category for a term. The presence of toggle words is a strong affect and triggered a switch of the concept's category. When toggle words are absent and more than one category matches, the disambiguation is more difficult. For our general

disambiguation we use the section type, the length of the lexicon match, the part of speech of the terms, and the presence of indicator phrases.

A few especially difficult and common terms were handled as special cases, for instance, term “PE”. This term most commonly means physical exam, however it can also mean pulmonary embolism or be part of the term “pe ct” which is a test for the same. For this case, if the term was “PE:” or near “admission”, “discharge” or “ct” we consider it a test; otherwise we call it a problem.

Results

Training set results

Number of Reference Files Tested :349.0

Number of Reference Lines :27837.0

Number of System Files Tested :349.0

Number of System Lines :26651.0

TESTING 1.1 - Exact span for all concepts together

	TP	FN	FP	R Value	P Value	F Value
Concept	21689	6148	4962	0.779	0.814	0.796
Class	21064	6773	5587	0.757	0.790	0.773

TESTING 1.2 - Exact span for separate concept classes

	TP	FN	FP	R Value	P Value	F Value
Problem	9025	2943	2340	0.754	0.794	0.774
Treatment	6788	1712	1420	0.799	0.827	0.813
Test	5876	1493	1202	0.797	0.830	0.813
Matching Class for Problem	8847	3121	2617	0.739	0.772	0.755
Matching Class for Treatment	6529	1971	1540	0.768	0.809	0.788
Matching Class for Test	5688	1681	1430	0.772	0.799	0.785

TESTING 1.3 - Inexact span for all concepts together

	TP	FN	FP	R Value	P Value	F Value
Concept	24821	3016	1830	0.892	0.931	0.911
Class	24821	3016	1830	0.888	0.901	0.894

TESTING 1.4 - Inexact span for separate concept classes

	TP	FN	FP	R Value	P Value	F Value
Problem	10648	1320	719	0.890	0.937	0.913
Treatment	7678	822	567	0.903	0.931	0.917
Test	6495	874	544	0.881	0.923	0.902
Matching Class for Problem	10398	1570	1066	0.869	0.907	0.888
Matching Class for Treatment	7339	1161	730	0.863	0.910	0.886
Matching Class for Test	6263	1106	855	0.850	0.880	0.865

The above results refer to our performance on the training set. TP refers to True Positives, FN to False Negatives, and FP to False Positives. R refers to Recall (sensitivity), P to Precision (specificity). F value is a statistical blend of the two. For R, P and F Values, higher is considered better. There are two main categories of results in the above

table: Exact and Inexact span matches. Given the wide variety of ways that a concepts can be expressed linguistically, exact span matches are much harder to obtain than inexact span matches. This fact is reflected in the above table, where our framework consistently does better in producing inexact matches (partial matches) as opposed to exact matches. Particularly, it seems that it's hardest to obtain the exact problem span, while at the same time designating it correctly as a problem.

Testing data results

Number of Reference Files Tested :477.0

Number of Reference Lines :45009.0

Number of System Files Tested :477.0

Number of System Lines :41410.0

TESTING 1.1 - Exact span for all concepts together

	TP	FN	FP	R Value	P Value	F Value
Concept	33049	11960	8361	0.734	0.798	0.765
Class	31839	13170	9571	0.707	0.769	0.737

TESTING 1.2 - Exact span for separate concept classes

	TP	FN	FP	R Value	P Value	F Value
Problem	13176	5374	3942	0.710	0.770	0.739
Treatment	10335	3225	2212	0.762	0.824	0.792
Test	9538	3361	2207	0.739	0.812	0.774
Matching Class for Problem	12906	5644	4470	0.696	0.743	0.718
Matching Class for Treatment	9882	3678	2556	0.729	0.795	0.760
Matching Class for Test	9051	3848	2545	0.702	0.781	0.739

TESTING 1.3 - Inexact span for all concepts together

	TP	FN	FP	R Value	P Value	F Value
Concept	38086	6923	3324	0.846	0.920	0.881
Class	38086	6923	3324	0.841	0.881	0.860

TESTING 1.4 - Inexact span for separate concept classes

	TP	**FN**	**FP**	R Value	P Value	F Value
Problem	15779	2771	1346	0.851	0.921	0.885
Treatment	11542	2018	1000	0.851	0.920	0.884
Test	10765	2134	978	0.835	0.917	0.874
Matching Class for Problem	15345	3205	2031	0.827	0.883	0.854
Matching Class for Treatment	10993	2567	1445	0.811	0.884	0.846
Matching Class for Test	10145	2754	1451	0.786	0.875	0.828

The Table above represents the performance results on the test set released on July 25th 2010. These results seem to mirror the results obtained from the training set, with a slight and expected reduction in performance overall. We say expected as the test set was not only larger than the train set, but it was also from a different source than we had seen in the train set. The latter exposed our method's slight overfitting to the training set, which was unavoidable considering that our performance was tuned using the training set. However, the reduction in performance was less than expected, measuring between 1% and 7% depending on the statistic being considered. Across all of the measurements, the overall average loss in performance was about 3.2%.

Conclusion

Overall, the results that our framework achieved on the training data had an F-measure for exact class span of 77.3%, inexact class span of 89.4%. The F-measures for the same two categories for the 477 test documents were: 73.4% & 86%.

The framework and approach used to detect concepts is being incorporated in practical applications which includes as one of the steps, the human review of annotated concepts. These efforts show much promise to improve the structuring of real-world clinical documentation.

Comment [Chuck Mul2]: do we know?

Comment [Oussama E3]: Reply to Chuck Mullett (08/30/2010, 10:08): "..."
We won't know until the second week of November or whereabouts.

Acknowledgements

We would like to thank the organizers of the NLP challenge to provide us the opportunity to work with marked up training data. Such corpus has been sorely lacking in this discipline. We also would like to thank our sponsor for this project, Kevin Piltz for allowing us to participate and work on this interesting challenge.

References

1. Uzuner, O., Luo, Y., Szolovits, P., " Evaluating the State-of-the-Art in Automatic De-identification," Journal of American Medical Informatics Association, Volume 14, Number 5, September/October 2007, pp 550-563.
2. Clark, C., Good, K., et al. "Identifying Smokers with a Medical Extraction System," Journal of Medical Informatics Association, Volume 15, Number 1, Jan/Feb 2008, pp 26-39.
3. Uzuner, O, Goldstein,I., Luo, Y., Kohane, I., "Identifying Patient Smoking Status from Medical Discharge Records", Journal of American Medical Informatics Association, Volume 15, Number 1, Jan/Feb 2008, pp 14-24.
4. Uzuner, Ö. (2009). "Recognizing Obesity and Co-morbidities in Sparse Data". *Journal of the American Medical Informatics Association.* July 2009; 16(4).
5. Ware, H., Mullett, J. C., Jagannathan, V., "Natural Language Processing (NLP) Framework to assess clinical conditions", to appear in the July 2009 issue of Journal of American Medical Informatics Association.
6. Alexander Turchin, Laura Morin, Luwam G. Semerec, Vipul Kashyap, Matvey B. Palchuk, Maria Shubina, Frank Changa, Qi L"Comparative Evaluation of Accuracy of Extraction of Medication Information from Narrative Physician Notes by Commercial and Academic Natural Language Processing Software Packages," Proceedings of AMIA 2006, pages 789-793.
7. Jagannathan, V., Mullett, CJ., et al. "Assessment of commercial NLP engines for medication information extraction from dictated clinical notes," Int J Med Inform. 2009 Apr;78(4):284-91.
8. Distributed Terminology System. [Online]. 2006 Available from: URL:http://www.apelon.com/products/white%20papers/DTS%20White%20Paper%20V34.pdf
9. Logical Observation Identifiers Names and Codes (LOINC) - http://loinc.org/. Accessed: 6/10/2010.
10. RxNorm. http://www.nlm.nih.gov/research/umls/rxnorm/ [Using it via: Apelon Utilites]
11. cTakes – From "OHNLP Documentation and Downloads" - https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/OHNLP_Documentation_and_Downloads Accessed: 6/1/2010
12. USHIK – United States Health Information Knowledgebase - http://ushik.ahrq.gov/index.html?Referer=Index Accessed: 6/20/2010
13.