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

Text Analysis Integration into a Medical Information Retrieval System: Challenges Related to Word Sense Disambiguation

A. R. Coden[1], G. K. Savova[2], J. D. Buntrock[2], I. L. Sominsky[1], P. V, Ogren[2], C, G. Chute[2] and P. C. de Groen[2]

[1]IBM, T.J. Research Center, Hawthorne, New York
[2]Division of Biomedical Informatics, Mayo Clinic, Rochester, Minnesota

Abstract

We describe a high throughput, real-time modularized text analysis system integrated into a production medical information retrieval system at the Mayo Clinic. The system identifies clinically relevant entities and concepts from unstructured medical reports, e.g. drugs, diagnosis and concepts, as defined in RxNorm, SNOMED CT and UMLS. To achieve the desired performance, the text analysis components were either developed for or adapted to the clinical domain. In this paper, we discuss one of several required innovations, disambiguation of mapped concepts with respect to an ontology. The methodology for the identification of a set of problem ambiguities and the creation of a manually sense-tagged data set for each of them is presented. Based on this data set, methods for word sense disambiguation for clinical notes were developed. In this paper, we report one set of F-scores (range is 0.47-0.98) which are substantially above the majority sense baseline.

Keywords:

natural language processing, word sense disambiguation, UMLS, UIMA

Introduction

With recent advances in Natural Language Processing (NLP) technology, the biomedical community has begun processing large amounts of textual unstructured data for a number of tasks. The MedTAS system developed by IBM is a modular text analysis and search system for the Life Sciences domain based on IBM's Unstructured Information Management Architecture (UIMA) ([1], [2]), http://incubator.apache.org/uima]. A source of errors in any NLP-based system is the mapping of a word token to multiple concepts ([3], [4], [5]). For example, "ms" maps to 10 UMLS concepts in the test set we created. Concept-based indexing and retrieval involves the unique mapping to a single representation. This is the problem of word sense disambiguation (WSD). In this paper, we present the design and architecture of a large-scale, highly modularized, real-time-enabled text analysis system and the development of its WSD component..

Methods

Figure 1 shows the architecture for bulk and real-time processing of clinical documents for subsequent information retrieval (IR), in production since 2005.

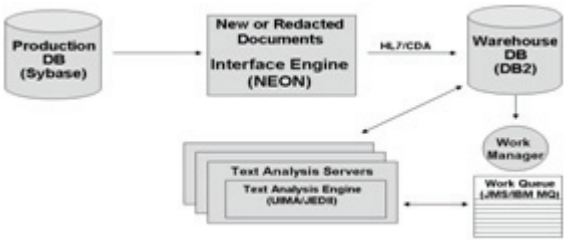

Figure 1 - Text analysis process flow

The UIMA-based text analysis system is a sequence of "annotators", each performing specific tasks. The annotators are native UIMA annotators, GATE-based and Open Source modules, and are either finite state automata or trained on hand-annotated corpora with a variety of machine learning (ML) algorithms. In addition to executing traditional NLP tasks on un-grammatical medical text, the annotators identify and extract named entities such as drugs, diagnosis and sign and symptoms allowing for concept-based indexing and retrieval. To increase retrieval results precision, multi-sense terms are to be uniquely mapped to a concept. Although there is a publicly-available WSD corpus derived from MedLine abstracts and manually sense-tagged with UMLS concepts ([5]), no comparable corpus for clinical ambiguities exists. We pose two specific questions: 1) are the most critical ambiguous terms in MedLine the same as those in the clinical domain, and 2) can the same algorithms be applied to both domains (biomedical literature and clinical text). Our first task was to create such a corpus.

The first step in the process was the identification of 50 frequent and highly relevant ambiguities from a medical retrieval perspective. We processed the entire 2002 collection of clinical notes through MetaMap ([6]), took the terms that mapped to multiple concepts and ranked them by their raw frequency. The top 1000 most frequent ambiguities were presented to four medical retrieval experts

who filtered them by relevancy for medical retrieval. The top 50 most relevant terms were included in our final WSD data set (referred to as Mayo WSD set). The corpus was then manually sense-tagged by four experts against UMLS. A "none of the above" category was added for those instances that did not have any sense representation in the UMLS.

The Mayo WSD set poses a number of challenges for any WSD annotator regardless of the technique employed. Some terms have over ten senses, e.g. "ac", thinly distributed over the 100 instances. Other senses have high representations of one specific sense, e.g. "ms". In addition, the "none of the above" category does not represent one monolithic sense.

Results

We used a ML approach to build the WSD annotator [7]. The experiments were based on 28 different feature sets on three variations of the corpus. The baseline F-score represented by the averaged majority sense is 0.6032. All feature sets performed well above the baseline. Detailed and comprehensive results will be presented in a subsequent paper. Here, we report one a set of F-scores, which was achieved by using a simple feature set of bag of unique stems filtered with a tf*idf formula within a window of 10 nonstop list tokens on both sides. This resulted in an average F-score of 0.776, well above the baseline.

Discussion

There are 9 ambiguous terms that overlap with the NLM WSD set described in [5]. Of note is that the senses themselves do not necessarily overlap. For almost all ambiguities, there are more senses in the clinical notes than in the literature abstracts, exemplifying that, in general, clinical notes are more complex than medical literature. Columns 4, 5 and 6 of Table 1 present the best results from three studies – [3], [4] and the current study. The first two studies worked with the NLM WSD set but excluded some ambiguities mainly because of the sparsity of the training instances. We experimented with the Mayo WSD set and achieved our best results with different feature sets than the one used to compute the F-score reported in the last column of Table 1. Extensive comparisons and error analysis will be presented in a subsequent publication.

Table 1 - Overlapping ambiguities between the Mayo WSD and NLM WSD set: number of senses in each set and best results from three studies

Ambiguity	Number of senses in Mayo WSD set (clinical notes)	Number of senses in NLM WSD set (literature abstracts)	Best F-score on NLM WSD set: Liu, Teller and Friedman	Best F-score on NLM WSD set: Leroy and Rindflesch	Best F-score on Mayo WSD set: current study
adjustment	5	4	not included in study	0.620	0.930
cold	6	6	0.909	not included in study	0.720
condition	6	3	not included in study	not included in study	0.900
discharge	2	3	0.908	not included in study	0.950
fit	5	3	not included in study	not included in study	0.740
lead	5	3	0.910	not included in study	0.900
sensitivity	7	4	not included in study	0.700	0.780
strain	9	3	not included in study	not included in study	0.690
support	6	3	not included in study	not included in study	0.740

Conclusion

We presented the Mayo Clinic IR system and part of our effort to incorporate a WSD component. For that, we created a manually sense-tagged set of 50 frequent and highly relevant ambiguities derived from the Mayo Clinic medical records, which is used to built and test a WSD annotator. In the future, we are planning to evaluate whether the disambiguation component indeed contributes to improved retrievals.

References

[1] Mack R, Mukherjaa S, Soffer A, et al. Text analytics for life science using the Unstructured Information Management Architecture. *IBM Systems Journal.* 2004;48(3):490.

[2] Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering.* 2004;10(3:4):327-348.

[3] Liu H, Teller V and Friedman C. A Multi-aspect Comparison Study of Supervised Word Sense Disambiguation. JAMIA, 2004, vol. 11, num 4, July/Aug.

[4] Leroy G and Rindflesch T. Using Symbolic Knowledge in the UMLS to Disambiguate Words in Small Datasets with a Naïve Bayes Classifier. *Proc Medinfo.2004*

[5] Weeber M, Mork JG, Aronson AR. Developing a test collection for biomedical word sense disambiguation. *Proc AMIA Symp.* 2001.

[6] Aronson AR, Bodenreider O, Chang HF, et al. The NLM Indexing Initiative. *Proc AMIA Symp.* 2000:17-21.

[7] Zhang T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *ICML'04*, pp. 919–926