Applying the TARSQI Toolkit to augment text mining of EHRs

Amber Stubbs
Department of Computer Science
Brandeis University MS 018
Waltham, Massachusetts, 02454 USA
astubbs@cs.brandeis.edu

Benjamin Harshfield
Channing Laboratory
Brigham and Women's Hospital
Boston, Massachusetts, 02115 USA
rebjh@channing.harvard.edu

Abstract

We present a preliminary attempt to apply the TARSQI Toolkit to the medical domain, specifically electronic health records, for use in answering temporally motivated questions.

1 Introduction

Electronic Health Records are often the most complete records of a patient's hospital stay, making them invaluable for retrospective cohort studies. However, the free text nature of these documents makes it difficult to extract complex information such as the relative timing of conditions or procedures. While there have been recent successes in this endeavor (Irvine et al., 2008; Mowery et al., 2009; Zhou et al., 2007), there is still much to be done. We present work done to adapt the TARSQI Toolkit (TTK) to the medical domain. Though the use of the TTK and a set of auxiliary Perl scripts, we perform information extraction over a set of 354 discharge summaries used in the R3i REALIST study to answer the following question:

> Which patients can be positively identified as being on statins at the time they were admitted to the hospital?

2 TARSQI Toolkit

The TARSQI Toolkit, developed as a part of the AQUAINT workshops, is a "modular system for automatic temporal and event annotation of natural language" in newswire texts (Verhagen and Pustejovsky, 2008). The different modules preprocess the data, label events and times, create links between times and events (called "tlinks"), and mark subordination relationships. Output from the TTK consists documents annotated in TimeML, an XML specification for event and time annotation (Pustejovsky et al., 2005). Of particular interest for this project are EVITA, the module responsible for finding events in text, and Blinker, the module used to create syntactic rule-based links between events and timexes.

3 Structure of EHRs

The bodies of the Electronic Health Records used were segmented, with each section having a header indicating the topic of that section ("Medical History", "Course of Treatment", "Discharge Medications", etc). Header names and sections are not standardized across EHRs, but often give important temporal information about when events described took place (Denny et al., 2008).

4 Statin Extraction Methodology

As the purpose of this task was to discover what changes to the TTK would be necessary to make the transition from newswire to medical texts, over the course of two weeks we filled in the gaps in the toolkit's abilities with a few auxiliary Perl scripts. Specifically, these scripts were used to clean up input so that it conformed to TTK expectations, label the statins as events, locate section headers and associate temporal information with the headers.

A list of statins was acquired from an MD, and then supplemented with information from websites in order to get all currently marketed versions of the drugs. This list was then used in conjunction with a Perl script to find mentions of statins in the discharge summaries and create TimeML event tags for them.

In order to identify and categorize section headers we developed a program to automatically collect header names from a separate set of approximately 700 discharge summaries. Then we gathered statistics on word frequency and created simple rules for characterizing headers based on keywords. Headers were divided into four simple categories: Past, Present, After, and Not (for cate-

Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 141–143,
Uppsala, Sweden, 15 July 2010.

gories that did not contain specific or relevant temporal information).

The Blinker component of the TTK was then modified to take into account temporal information stored in the header in addition to the syntactic information present in each individual sentence for the creation of tlinks.

5 Results

Output from the modified TTK was compared to the judgment of human annotators on the same dataset. Two annotators, employees of BWH/Harvard Medical involved in data management and review for clinical trials, were asked to label each file as yes for those patients taking statins at the time they were admitted to the hospital, and no for those that werent. Files where statins were mentioned without clear temporal anchorings were categorized as "unsure".

Inter-annotator agreement was 85% (Cohen kappa=.75), with 75% of the disagreements being between "no" and "unsure". The majority of these ambiguous cases were discharge summaries where a statin was listed under "discharge" but admission medications were not listed, nor were the statins mentioned as being started at the hospital. The annotation guidelines have been updated to reflect how to annotate these cases in the future. Overall, 139 patients were identified as being on statins, 174 were not on statins, and 41 were unclear.

As the question was which patients could be positively identified as being on statins at the time of admission, the files labeled as "unsure" were considered to be "no" for the purposes of evaluation against the TTK, making the totals 139 yeses to 215 noes. The comparison between human and computer annotation are shown below:

	Yes	**No**
Human	139	215
TTK	129	225

Table 1: Distribution of statin classifications.

The TTK system had an accuracy of 84% overall, with an accuracy of 95% on the files that the human annotators found to be unambiguous.

6 Limitations

While we were pleased by these results, a number of factors worked in the favor or the automated system. The task itself, while requiring a mixture of lexical and temporal knowledge, was greatly simplified by a finite list of medications and a binary outcome variable. Obscure abbreviations or misspellings could have prevented identification of statin mentions for both the computer and humans, making the overall accuracy questionable. Additionally, in the majority of documents the statins were mentioned in lists under temporally anchored headings rather than free text, thereby minimizing the impact of uncertain times as described in Hripcsak et al (2009).

7 Future work

Our work so far shows promising results for being able to modify the TARSQI Toolkit for use in the medical domain. In the future, we would like to integrate the functionality of the Perl scripts used in this project into the TTK, in particular expanding the vocabulary of the EVITA module to the medical domain, section header labeling, and the use of the headers in tlink creation.

New annotation schemas will need to be added to the project in order to get a more complete and accurate view of medical records. Under consideration is the Clinical E-Science Framework (CLEF) (Roberts et al., 2007) for annotating medical entities, actions (which would overlap with TimeML events), drugs, etc. Modifications to Blinker will be more fully integrated with the existing rule libraries. At this point it is unclear whether the TTK will remain a single program, or if it will split into domain-specific versions.

Furthermore, the number of files labeled "unsure" by human annotators highlights the need for cross-document analysis abilities. Had previous records for these patients been available, it seems likely that there would have been fewer uncertainties.

8 Conclusion

Modifying the TARSQI Toolkit, a newswire-trained parser, for application in the medical domain provided accurate results for a very specific time-sensitive query.

Acknowledgments

Partial support for the work described here was provided by the Residual Risk Reduction Initiative Foundation (r3i.org).

References

Joshua C Denny, Randolph A Miller, Kevin B Johnson, and Anderson Spickard. 2008. Development and evaluation of a clinical note section header terminology. *AMIA Annual Symposium proceedings*, pages 156–60.

George Hripcsak, Noémie Elhadad, Yueh-Hsia Chen, Li Zhou, and Frances P Morrison. 2009. Using empiric semantic correlation to interpret temporal assertions in clinical texts. *J Am Med Inform Assoc*, 16(2):220–7.

Ann K Irvine, Stephanie W Haas, and Tessa Sullivan. 2008. Tn-ties: A system for extracting temporal information from emergency department triage notes. *AMIA Annual Symposium proceedings*, pages 328–32.

Danielle L. Mowery, Henk Harkema, John N. Dowling, Jonathan L. Lustgarten, and Wendy W. Chapman. 2009. Distinguishing historical from current problems in clinical reports: which textual features help? In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 10–18, Morristown, NJ, USA. Association for Computational Linguistics.

James Pustejovsky, Bob Ingria, and Roser Sauri et al., 2005. *The Language of Time: A Reader*, chapter The Specification Language TimeML, pages 545–558. Oxford University Press, Oxford.

Angus Roberts, Robert Gaizauskas, and Mark et al Hepple. 2007. The clef corpus: semantic annotation of clinical text. *AMIA Annual Symposium proceedings*, pages 625–9.

Marc Verhagen and James Pustejovsky. 2008. Temporal processing with the tarsqi toolkit. In *Coling 2008: Companion volume - Posters and Demonstrations*, pages 189–192, Manchester, UK.

Li Zhou, Simon Parsons, and George Hripcsak. 2007. The evaluation of a temporal reasoning system in processing clinical discharge summaries. *J Am Med Inform Assoc*, 15(1):99–106.