Application of Natural Language Parsers To Syndromic Surveillance

Michael J. Waddell, Ph.D., César R. Meraz, M.A.
Pangaea Information Technologies, Chicago IL
Julio C. Silva, M.D.,M.P.H, Dino P. Rumoro, D.O.
Rush University Medical Center, Chicago IL

Objective

This paper describes a methodology for applying natural language parsing (NLP) technologies, originally developed for analyzing biomedical journal articles, to the monitoring of emergency department patient charts for infectious diseases of interest.

Background

Rush University Medical Center and Pangaea Information Technologies are currently engaged in a multi-year collaboration to develop GUARDIAN™, a syndromic surveillance system that monitors emergency department (ED) patient data for diseases of interest[1,2]. An important part of this system is a tool that mines free text for words/phrases associated with predefined disease profiles. Currently, this uses MetaMap Transfer (MMTx), a software component developed at the National Library of Medicine, to decompose text into sentences, phrases, and concepts.

To overcome a limitation of MMTx – namely lack of negation detection – GUARDIAN™ also uses NegEx[6]. The NegEx algorithm determines – for each concept returned by MMTx – if the context in which it is found *asserts* that concept or *denies* it. Using this tool to discover concepts expressed in the text provides more accurate access to information stored in free-text fields such as a history of present illness (HPI). This allows GUARDIAN's inference engine to better quantify the "proximity" of a patient's particular set of symptoms to a given disease profile.

Both MMTx and NegEx were developed for parsing the "well-formed" language found in biomedical journal articles. The combination of both of these technologies has been successfully employed and tested in prior studies using medical journal articles[3,4,5,6]; however, patient charts lack the same editorial process and thus commonly exhibit poor grammar and spelling. MMTx and NegEx must be adapted in order for them to be successfully used with this highly *un*structured text. The motivation behind this research is to determine, quantify and validate how to best perform these adaptations.

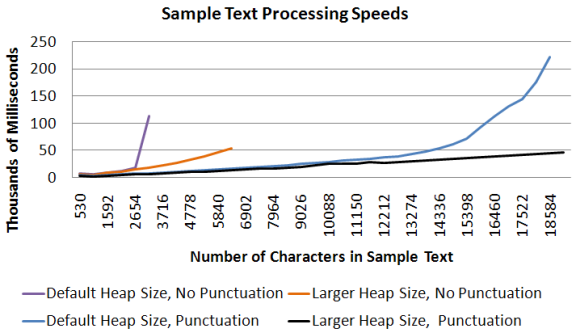
Methods

First, we itemized each significant way, from the perspective of this task, that ED medical charts differ from biomedical journal text; and identified how each adversely affects the use and/or performance of MMTx and/or NegEx. Second, we quantified the impact of each of these effects. Third, we proposed ways to minimize or eliminate the most significant impacts. Finally, we tested our proposed changes to validate how successfully we addressed these issues.

All of our tests used a sample HPI taken from an actual patient chart. To test changes in sample length, without changing the relative composition of the sample, we appended multiple copies of this sample to itself. Early in our testing, we increased Java's memory allocation (heap size) setting to 512MB in order to prevent excessively long run-times.

Results

The single most significant difference between ED medical charts and biomedical journal text is the poor grammar and spelling found in chart data[†]. Without punctuation, MMTx will consider all input as a single sentence. This dramatically increases run-times and will cause the system to fail completely when a single sentence contains more than 1000 words (see below).



Conclusions

Increasing memory allocation dramatically improved system performance, but using standard punctuation delivers both performance gains and improved sentence decomposition. Because ED patient charts often do not contain reliable punctuation, we propose a novel algorithm – which incorporates the work of Liu, et al.[7] – in order to address this limitation. Using this in a "pre-parser" allows both MMTx and NegEx to process and apply their results to proper sentences.

References

[1] Waddell, M; et al. *Poster*. Syndromic Surveillance 2007.
[2] Silva, J; et al. *Poster*. Syndromic Surveillance 2007.
[3] Johnson, S; et al. JAMIA. 2008;15:54-64.
[4] Friedman, C; et al. Bioinformatics, 17 Suppl 1:S74–82, 2001.
[5] Rzhetsky, A; et al. Bioinformatics, 16(12):1120–8, 2000.
[6] Chapman, W; et al. AMIA Symposium 2001, pp 105–9.
[7] Liu, Y; et al. EMNLP 2004.

[†] The full list of differences is beyond the scope of this abstract.