Statistical Extraction of Medication Information from Clinical Records

Scott Russell Halgrim[1], MA, Fei Xia[1], PhD, Imre Solti[1], MD, PhD,
Eithon Cadag[1], PhD, Ozlem Uzuner[2,3], PhD
[1]University of Washington, Seattle, WA; [2]University of Albany, SUNY, Albany, NY;
[3]Middle East Technical University, Northern Cyprus Campus, Mersin 10, Turkey

Abstract

We propose a statistical medication extraction system and show that it significantly benefits from contextual features.[‡]

Introduction

Clinical notes often represent medication information as free text; automatic extraction of this information can positively impact clinical data management[1]. The 2009 i2b2 challenge (https://www.i2b2.org/NLP/Medication/) focused on filling medication *entry* templates with six *named entities* (medication name, dosage, frequency, duration, mode and indication), and evaluated 20 teams from around the world. 696 raw discharge summaries were released, 17 of which were annotated. An additional 251 discharge summaries annotated by the community were the test set. We trained our statistical medication extraction system on 145 annotated discharge summaries courtesy of University of Sydney and evaluated it on the challenge test set.

Methods

We present a maximum entropy classifier that labels each word in a document according to a *BIO* scheme (*e.g.*, the tag *B-mode* means the word is the beginning word in a dosage, *I-mode* means it is inside a mode, and O means it is outside all named entities)[2, 3]. The tag sequences identify named entities. For instance, the sequence "w_1*/B-name* w_2*/I-name* w_3*/O* w_4*/B-mode*" indicates the text contains two spans: "w_1 w_2", a medication name, and "w_4", a mode.

Experimental Results

Table 1 shows *exact* and *inexact horizontal system-level* F-scores of extracted medication entries. Results yield increasing performance with the incremental addition of feature sets: (1) word unigrams, (2) word bigrams/trigrams and selected word properties (*e.g.*, affixes), and (3) labels of previous words and features beyond *n*-grams and spelling[3]. We refer to sets (2) and (3) as contextual features.

	(1)	(1)+(2)	(1)+(2)+(3)
Exact	57.1	79.3	84.1
Inexact	59.4	78.0	83.9

Table 1: System F-scores: step-wise differences between adjacent columns are statistically significant (by approximate randomization). For comparison, the top ten 2009 i2b2 challenge systems ranged from 76.4 to 85.7 for exact, with an average of 79.7. Inexact scores ranged from 75.9 to 84.9 with an average of 79.8[4].

Conclusion

The results show that a statistical approach works well for the task and that contextual features bring significant improvement.

Acknowledgments

This work was supported in part by NIH Grant U54LM008748 and grants T15LM007442-06 and 1K99LM010227-0110.

References

1. Levin MA, Krol M, Doshi AM, Reich DL. "Extraction and mapping of drug names from free text to a standardized nomenclature." AMIA Annu Symp Proc. 2007 Oct 11:438-42.
2. Saha SK, Sarkar S, Mitra P. "Feature selection techniques for maximum entropy based biomedical named entity recognition." *Journal of Biomedical Informatics*. 2009 October; 42(5): 905-11.
3. Halgrim S. *A Pipeline Machine Learning Approach to Biomedical Information Extraction.* Thesis. University of Washington, 2009.
4. Uzuner O, Solti I, Cadag E. "i2b2 NLP Challenges." *Third i2b2 Workshop: Challenges in Natural Language Processing for Clinical Data Medication Extraction Challenge.* San Francisco. 13 Nov. 2009.

[‡] As workshop co-organizers with extended access to the data, we did not submit this system to the challenge. To avoid ranking ourselves against the systems that participated in the challenge, we compare our results against the mean performance of the top ten systems rather than individual challenge systems.