Addressing the Annotation Bottleneck for Clinical Natural Language Processing: Testing the Feasibility of Domain Adaptation for Medical Text

Journal:	*AMIA 2010 Annual Symposium*
Manuscript ID:	AMIA-1889-A2009
Manuscript Type:	Poster
Date Submitted by the Author:	16-Mar-2010
Complete List of Authors:	Solti, Imre; University of Washington, Biomedical and Health Informatics Halgrim, Scott Xia, Fei; University of Washington, Department of Linguistics

Addressing the Annotation Bottleneck for Clinical Natural Language Processing: Testing the Feasibility of Domain Adaptation for Medical Text

Imre Solti, MD, PhD, Scott R. Halgrim, MA, Fei Xia, PhD
University of Washington, Seattle, WA

Abstract

Privacy regulations created a serious bottleneck for accessing and annotating text in EMR. Statistical natural language processing requires annotated training text. There are publicly available biomedical text sources. We propose to apply domain adaptation to (source) clinical trial announcements and (target) physician discharge summaries (DS) to ease the annotation bottleneck. We tested Augment and InstancePruning. Augment improved the performance of medication name extraction statistically significantly on p=0.0006 level when only few annotated DS were available.

Introduction

The Health Insurance Portability and Accountability Act (HIPAA) of 1996 presented a set of federal standards for protecting the privacy of personal health information and it seriously limited data access in electronic medical records[1]. Statistical Natural Language Processing (NLP) requires annotated training data, consequently HIPAA created a bottleneck for statistical NLP research. We propose to apply domain adaptation to ease the annotation bottleneck. In this poster we report our preliminary findings and discuss future research directions.

Related Work

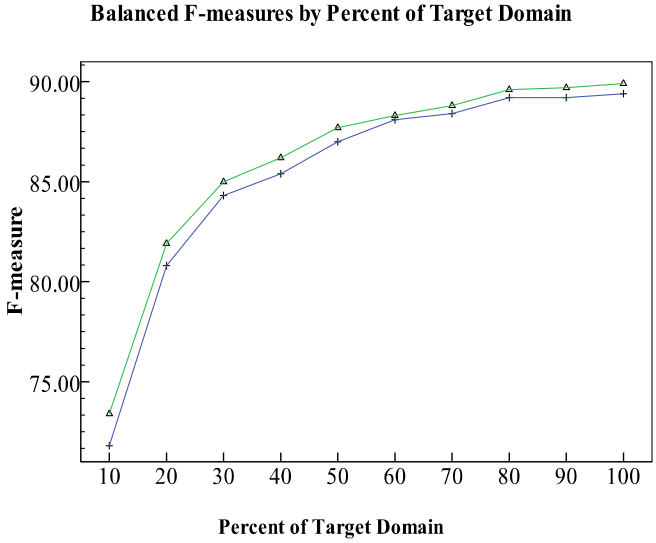
There are supervised (requires labeled target domain text) and unsupervised (no labeled target text) adaptations. Our task falls into the first category. Jiang (2008) discusses two frameworks for adaptation: instance weighting and feature selection[2].

Data and Methods

In this experiment, we identify medication names in biomedical text. The source domain (no restrictions) currently includes 100 randomly selected Clinical Trial Announcements. The target domain (strict HIPAA restrictions) includes 110 discharge summaries for training, 35 for development, and 251 for testing. We tested a Maximum Entropy algorithm on the source and target domains using InstancePruning and Augment adaptation algorithms[3].

Results and Discussion

We calculated three baselines: (1) source data only, (2) target data only, (3) use both source and target data. (Figure 1) shows the results of the target only baseline and Augmentation algorithms. Augmentation has higher numerical value for all experiments and the difference is statistically significant at p=0.0006 level.



Conclusion

We proved that domain adaptation from publicly accessible clinical trial announcements to physician discharge summaries is feasible and provides statistically significant improvement for medication name extraction. In the future we plan to quantitatively measure the domain difference between source and target domains, and use that finding to guide our choice of adaptation algorithms

Acknowledgment

1K99LM010227-0110.

References

1. Committee on Health Research and the Privacy of Health Information. Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. Washington, DC: Institute of Medicine; 2009.
2. Jiang, J. Domain adaptation in natural language processing. Doctoral Dissertation UMI # 3337811
3. Daume III, H. Frustratingly easy domain adaptation. Proc 45th Ann Meet ACL, 256-263, 2007.