Bypassing Knowledge Acquisition Bottleneck with Bayesian Word Sense Induction

Anna Rumshisky
Department of Computer Science
UMass Lowell
Lowell, MA 01854
arum@cs.uml.edu

Rachel Chasin
CSAIL
Massachusetts Institute of Technology
Cambridge, MA 02139
rchasin@mit.edu

Peter Potash
Department of Computer Science
UMass Lowell
Lowell, MA 01854
ppotash@cs.uml.edu

Abstract

We use Bayesian topic modeling techniques adapted to the task of unsupervised word sense induction on acronyms in clinical text and investigate (1) the amount of annotated data needed by such approaches to match the performance of the supervised sense disambiguation systems, and (2) feasibility of using an automatically collected silver standard for such techniques. A dataset of ambiguous abbreviations from the University of Minnesota Duluth is used to examine the effect of the amount of labeled data on accuracy. On the abbreviation dataset, performance levels out at 60-70 labeled instances.

1 Introduction

Clinical narrative is one of the richest still mostly untapped sources of clinical information contained in Electronic Health Record (EHR). The text of provider notes is known to be noisy and often ungrammatical, brimming with domain-specific and often highly ambiguous abbreviations and acronyms, which often vary even across different institutions. The proliferation of ambiguity in clinical text is often a major roadblock to effective information extraction. However, it is still a largely unsolved problem, since applying tried and true supervised disambiguation methods transplanted from the general domain would require extensive expert annotation of a large amount of text for each target word.

We have recently shown that such unsupervised methods as topic modeling can be successfully adapted to the problem of word sense induction (WSI) in the clinical domain, and produce superior results for non-acronym related ambiguities [1]. Following Brody and Lapata [3], we treat each sentence containing an ambiguous word as a document, and the derived topics as sense-selecting context patterns represented as collections of features.

Topic models are trained on a large corpus of unlabeled data from the MIMIC II (Multiparameter Intelligent Monitoring in Intensive Care) database. The topic modeling method also requires some labeled data to create a mapping between the derived topics and the senses present in the gold standard. The gold standard is therefore split into a mapping set and an evaluation set. In this paper, we use the acronym data set developed by Moon et al. [2] as the gold standard to investigate the following questions:

(1) Data set size required by the unsupervised methods. How much data is required by such methods to obtain accurate mappings of derived sense clusters to senses? At what point does the performance level off? Do such methods require less data than supervised learning methods to obtain a similar performance?

(2) Silver standard sense-annotated data. Is automatically creating silver-standard sense-annotated data for acronyms feasible? For acronyms sense annotation amounts to acronym expansion. Automatically collecting a good silver-standard dataset mapping acronyms to expanded long forms is a reasonable task. We investigate the quality of results obtained with such a set.

2 Related Work

Compared to general WSD, abbreviation expansion is more conducive to semi-supervised approaches in which a silver-standard dataset is collected automatically, typically by searching for the long forms in a corpus and replacing them with their abbreviations [4, 5]. The two main issues that may arise in this method are the lack of long forms appearing in clinical text and the differences in contexts surrounding long forms when they do appear.

3 Acronym/Abbreviation Data

Evaluation Data. The labeled abbreviation data (henceforth, Abbr) consists of 75 ambiguous clinical abbreviations with 500 instances each [2] drawn primarily from dictated notes. Each instance contains a few sentences of context and a manually-assigned label (long form). For the 50 targets that we use, the most frequent sense (MFS) accounts for less than 95% of examples. We use mapping sets consisting of up to 120 instances, since according to Moon et al. [2], 125 instances provided sufficient training data for supervised methods. The remaining 380 instances per target are used for testing.

Training Data. Clinical text from the MIMIC II database was used to obtain unlabeled training data for the topic models. For each target acronym, up to 50,000 instances were extracted from the nursing notes and discharge summaries contained in the database, with 100-token context included for each instance.

4 Methods

4.1 Topic Modeling Methods

For each target word, a topic model is trained on a large unlabeled data set, and applied to the held-out labeled set, so that each example is associated with a distribution over senses. The labeled set is then used to map the derived set of senses to gold standard long forms associated with each example, following the supervised evaluation methodology used in the general domain for word sense induction tasks [6, 7]. We use a Bayesian topic modeling method, Latent Dirichlet Allocation (LDA) [8] and a non-parametric generative model Hierarchical Dirichlet Process (HDP) [9] that does not require a pre-defined number of topics. We train topic models using two kinds of features: (1) stemmed bag-of-words and (2) concept features derived from the Unified Medical Language System (UMLS) metathesaurus, using 6 closest words and 6 closest concepts as features. Clinical features are created as follows. Each string of up to six tokens is normalized using the Lexical Variant Generation program (LVG) which tokenizes, stems, and alphabetizes the resulting tokens [10]. The normalized string is looked up in the English normalized string table (mrxns_eng) provided by the UMLS and if present that string is considered a clinical concept.

4.2 Effects of Mapping Set Size

Since manually annotated data is costly, we would like to minimize the size of the mapping set without sacrificing accuracy. We experiment with the effect of this size on accuracy. From 500 instances per target in Abbr, we create mapping sets $k_1 \ldots k_{12}$ of sizes 10, 20, ..., 120 where set k_i is contained within set k_{i+1}. We pick one LDA and one HDP configuration that performed well in our previous experiments, using 6 topics per acronym for LDA [1]. We train models for these on the training set collected from MIMIC II database for Abbr targets. We then use each of the 12 sets

in turn as the mapping set, and test on the 380 instances not found in any of the mapping sets. As we are only interested in the effect of mapping set size on accuracy in these experiments, we do not further vary the model configurations.

4.3 Automatically Created Mapping Set

The topic-modeling approach to WSI does not require large amounts of mapping data, as seen in Section 5.3.1, but with vast numbers of possible ambiguous targets, automatic generation of this data would make the method much more scalable. Abbreviation expansion is unique in that, unlike for other kinds of lexical ambiguity, it is possible to obtain "silver standard" labeled corpora automatically. We experiment with this idea by creating a mapping set for the Abbr targets from notes in MIMIC. Following the methods of Xu et al. [4] and Pakhomov et al. [5], we look for occurrences of an abbreviation's long forms in MIMIC (using exact string matching), collect a small context around each, and replace the long form with the abbreviation, labeling the created instance with the long form. For example, "the patient remained on antibiotics, off pressors and on a ms drip for comfort" is part of an instance labeled 'morphine sulfate' while "The patient is a 41-year-old man with advanced ms who is paraplegic secondary to this" is part of one labeled 'multiple sclerosis'. We find the long forms for matching by looking up the abbreviations in LRABR, a table of the SPECIALIST Lexicon. To keep this task feasible, we collect up to 1000 instances per long form.

The long forms from LRABR that appeared in MIMIC are quite different from the ones in the Abbr labeled data. For each target, examining the forms with $> 1\%$ frequency yields only 21/75 targets with more than one form in common between the automatically collected set and the Abbr test set. Most of these common forms have very different frequency distributions in the two corpora, making it difficult to use Abbr for evaluation. We evaluate the accuracy for eight targets with higher overlap between senses in corpora: bm, cva, er, mr, ms, otc, pda, and ra.

5 Results

5.1 Mapping Set Size Experiments

Figure 1 shows the accuracies from the mapping set size experiments on Moon's 50 targets from the Abbr test set (380 instances/target). We use LDA 6w +6c with 6 topics and HDP 6w +6c configurations in these experiments. The HDP configuration performs better, and levels out around 70 instances; LDA levels out around 60. The accuracies get quite high, above 85%. While Moon et al. [4] have reported higher accuracy achieved by an RBF kernel Support Vector Machine classifier trained on the same amount of labeled data, however, we were unable to replicate these results. Thus, in our experiments, a similar classifier trained on 100 instances only achieves the accuracy of 68.1%. For comparison, we also show the test set MFS baseline, corresponding to the accuracy obtained if the majority sense is predicted for every instance.

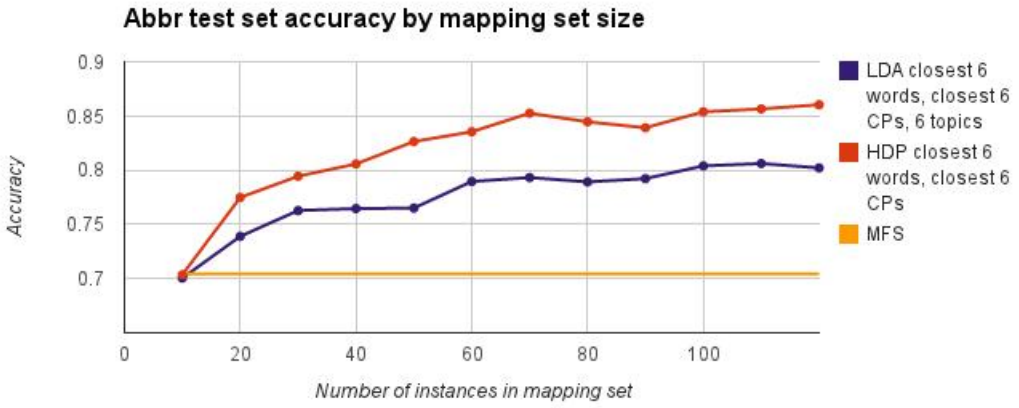

Figure 1: Accuracy on the Abbr test set as a function of the number of instances used for mapping.

5.2 "Silver" Mapping Set Experiments

Table 1 shows the results of the experiments with automatic mapping sets created for 8 targets. These results are contrasted with the results using the 70-instance mapping set from Abbr. Since there is a difference in genre of the notes in the training and test data (Abbr is drawn from dictated notes, while MIMIC data is not), it is not surprising that the Abbr mapping set gives somewhat higher performance on the test data. However, in a realistic application, disambiguation needs to be performed on the text similar to the one from which the mapping set is collected, making the automatic mapping set creation more feasible.

Target	**MFS**	**LDA, 6w +6c, 6 topics, auto-mapping**	**LDA, 6w +6c, 6 topics, *Abbr* mapping**
bm	89.7%	26.1%	89.7%
cva	58.7%	75.8%	96.6%
er	88.9%	93.2%	88.9%
mr	65.0%	94.2%	94.7%
ms	55.0%	85.8%	91.8%
otc	93.4%	93.4%	93.4%
pda	72.9%	90.8%	90.3%
ra	79.7%	65.0%	79.5%

Table 1: Results of using an automatic mapping set for eight targets and comparison to using a gold-standard mapping set

6 Discussion

Our experiments with abbreviations indicate that a relatively low amount of labeled data allows the mapping stage of the topic modeling method to maximize performance. However, the task of automatically collecting this data depends heavily on whether a particular long form is likely to occur in the text. While the "silver" mapping set experiments fail to match the accuracy obtained with the gold standard data, error analysis indicates that this is largely due to the difference in genres between the unlabeled data from which the silver standard was drawn and the evaluation data.

References

[1] Chasin, R., Rumshisky, A., Uzuner, O. & Szolovits, P. (2013) Word sense disambiguation in the clinical domain: A comparison of knowledge-rich and knowledge-poor unsupervised method. Submitted to *Journal of American Medical Informatics Association*

[2] Moon, S. Pakhomov, S. & Melton, G.B. (2012) Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations. *AIMA Annual Symposium Proceedings 2012*: 1310-9.

[3] Brody S. & Lapata M. (2009) Bayesian word sense induction. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*

[4] Xu, H., Stetson, P.D. & Friedman, C. (2012) Combining Corpus-derived sense profiles with estimated frequency information to disambiguate clinical abbreviations. *AIMA Annual Symposium Proceedings 2012*:1004-13.

[5] Pakhomov, S., Pederson, T. & Chute C.G. (2005) Abbreviation and acronym disambiguation in clinical discourse. *AMIA Annual Symposium Proceedings 2005*:589-593.

[6] Agirre E. & Soroa A. (2007) Semeval-2007 task 2: Evaluating word sense induction and discrimination systems. *Proceedings of the Fourth International Workshop on Semantic Evaluations*:7-12.

[7] Manandhar, S., Klapaftis, I.P., Dligach, D., et al. (2010) SemEval-2010 task 14: Word sense induction & disambiguation. *Proceedings of the 5th International Workshop on Semantic Evaluation*:63-68.

[8] Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003) Latent dirichlet allocation. *Journal of Machine Learning Research*;**3**:993-1022.

[9] Teh, Y.W., Jordan, M.I. , Beal, M.J. & Blei, D.M. (2006) Hierarchical dirichlet processes. *Journal of the American Statistical Association*;**101**(476):1566-1581.

[10] Bodenreider O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*;**32**(1):D267-D270.