BMC Bioinformatics

INTRODUCTION Open Access

Topics in machine learning for biomedical literature analysis and text retrieval

Rezarta Islamaj Doğan[†], Lana Yeganova[*†]

From Machine Learning for Biomedical Literature Analysis and Text Retrieval in the International Conference for Machine Learning and Applications 2010
Washington, DC, USA. 12-14 December 2010

Introduction

Life science researchers and health care professionals rely heavily on biomedical literature databases such as MEDLINE to access information essential for research, health care, education, as well as to keep up with the latest developments in their fields. Providing ways to efficiently access and analyze text information is critical and is becoming more challenging with the increasing volume of publications in the biomedical domain. The last decade has shown an exponential rate of growth of biomedical literature [1].

Natural language processing, a symbiosis of computer science and linguistics disciplines, addresses the computational aspects of automatic text processing. This field offers a fertile ground for machine learning algorithms. The challenges presented when processing natural language offer new opportunities to the existing machine learning methods and promote the development of new ones.

The special session of "Machine Learning in Biomedical Literature Analysis and Text Retrieval" was held for the first time as part of the 9th International Conference on Machine Learning and Applications, in Washington DC on December 12-14, 2010. The goal of this session was to present advancements in machine learning techniques that can improve the analysis of biomedical text.

In this supplement we present a collection of papers originally presented and published in the proceedings of the International Conference on Machine Learning and Applications (ICMLA 2010). These papers constitute an advance beyond the work originally presented at the conference and have gone through a separate rigorous review process. They represent a wide cross-section of the type of work that goes on in machine learning today, with its focus on biomedical literature.

Papers in this supplement touch on multiple existing machine learning methods such as wide margin classifiers and conditional random fields. They suggest novel applications for these methods as well as propose new machine learning techniques, such as novel methods for constructing training data and gold standards. From the literature analysis and text retrieval perspectives this collection of papers covers multiple topics including tokenization, named entity recognition, word-sense disambiguation, sequence labeling, and relationship extraction.

Tokenization is typically the first step in natural language processing and is often assumed to be trivial. Unfortunately, it is quite challenging, especially in the biomedical domain. Barrett and Weber-Jahnke [2] present an intriguing scheme for building a tokenizer.

Named entity recognition is an important component of text analysis tools. Three papers in the supplement touch on named entity recognition. Yeganova et al. [3] present a method of detecting abbreviations and their definitions in biomedical literature. Islamaj Dogan et al. [4] present an approach that detects with high accuracy clinical problems, treatment and test phrases in patient records and doctor notes. Benton at al. [5] present a system for de-identifying personal information in medical message board text.

Many applications are believed to benefit from identifying the correct word sense in entity recognition tasks. MetaMap [6], for example, is a system that provides UMLS [7] concept and semantic type annotation to free text and can significantly benefit from word-sense

* Correspondence: Yeganova@mail.nih.gov
† Contributed equally
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
Full list of author information is available at the end of the article

disambiguation. Jimeno-Yepes et al. [8] work on a knowledge-based *word sense disambiguation* approach that uses collocation analysis to improve the knowledge-based word sense disambiguation system.

Automatic extraction of bibliographic data, such as article titles, author names, abstracts, and references are essential to citation databases, such as MEDLINE. Zhang et al. [9] examine the task of identifying the components of bibliographic references. They treat the problem as a *sequence labeling problem*.

Accessibility to gold-standard *training data* allows scientist to focus on the solution of the problem at hand. In this collection we include two papers that are dedicated to this issue. Wilbur and Kim [10] treat human relevance judgments of MEDLINE document pairs to improve on gold standard annotations, whereas Yeganova et al. [3] present a method that relies on naturally occurring positive training examples and synthetically generated negative training examples to train their model.

Finally, Islamaj Dogan et al. [4] investigate a clinical *relationship extraction* problem. They approach it as a classification task, training classifiers to assign a relationship type to a pair of clinical concepts after performing entity recognition.

Summary of selected contributions

In this section we present a short summary for each of the contributions to this supplement.

Building a biomedical tokenizer using the token lattice design pattern and the adapted Viterbi algorithm, Barrett N, Weber-Jahnke J.

In their paper Barrett and Weber-Jahnke [2] present an intriguing scheme for building a tokenizer. The system consists of three components:

1) Token transducers that convert running text into a sequence of tokens and corresponding part of speech (POS) tags;

2) A token lattice constructor that constructs a bounded token lattice from the outputs of the different token transducers applied to a piece of text;

3) A best path selection algorithm (a variation of Viterbi algorithm) that selects an optimal sequence of tokens to represent the text.

The key idea of the study is to train an HMM on POS tagged training data and use the result with the Viterbi algorithm to assign the most likely tokenization of a text string as its' tokenization. Testing shows that results compare favourably with other biomedical taggers, which suggests that this is a good approach for creating biomedical text taggers.

A system for de-identifying medical message board text, Benton A, Hill S, Ungar L, Chung AP, Leonard C, Freeman C, Holmes JH.

There are millions of public posts to medical message boards by users seeking support and information on a wide range of medical conditions. This paper [5] describes a novel method to de-identify medical message board postings, in order to be able to use these postings for research purposes. Benton et al. develop and test a system to detect various kinds of Personally Identifiable Information. To obtain anonymized text, the system utilizes regular expressions and conditional random field classifiers. The novelty of the system is their approach to identify names by using their frequency distribution– their system is looking for words that occur at high frequency within a given document, but low frequency over all. Authors test their system on breast cancer and arthritis corpora and demonstrate that their approach outperforms publicly available systems.

A context-blocks model for identifying clinical relationships in patient records, Islamaj Doğan R, Névéol A, Lu Z

Patient records are document summaries that are given to patients after they are discharged from the hospital. In this paper Islamaj Dogan et al. [4] provide a system that identifies main clinical concepts such as medical problems and treatments, and their interdependent relationships, which is the preliminary step for many applications in medical informatics.

The detection of clinical problems, treatments and tests is treated as a sequence labeling problem. The authors use conditional random fields in combination with Priority Model [11] to accurately identify valuable concepts in clinical text. The detection of clinical relationships, such as treatment improves/worsens clinical problem, is abstracted as a form of five building blocks, each of which captures a different part of the context of the relationship. As such, this relationship schema can be easily visualized geometrically, and given a sentence where two concepts are present, such a sentence can be dissected into distinct segments. The authors present an end-to-end system, from concept recognition to relationship detection.

Collocation analysis for UMLS knowledge-based word sense disambiguation, Jimeno-Yepes A, McInnes BT, Aronson AR.

Word sense disambiguation is another important problem that may affect the accuracy of text analysis for information retrieval and text mining. In their paper, Jimeno-Yepes et al. [8] present an evaluation of a number of different methods to address the problem. They analyze collocation types which could improve the performance of the knowledge-based disambiguation method. They examine two methods for assigning collocations to target senses. They also measure and report the effect of these collocations on two Word Sense Disambiguation methods.

Improving a gold standard: treating human relevance judgments of MEDLINE document pairs, Wilbur WJ, Kim W.

Human relevance judgments of a document are important as a means of evaluating the performance and as a source of training data for machine learning methods. Because human judgments are difficult, time consuming and expensive to obtain, it is important to extract as much advantage or information from human judgments as possible. If one is fortunate enough to have multiple judgments for the reasonably large collection of similar objects, then it is possible to make predictions of future human judgments for the whole collection that are superior to the simple maximum likelihood estimate. This is possible because the multiple judgments over the collection allow determining the relative value of a judge as compared with the other judges in the group and this value can be used to augment or diminish a particular judge's influence in predicting future judgments. In their paper, Wilbur and Kim [10] study and compare five different methods for estimating the labeling probability distribution and show that each is superior to simple maximal likelihood estimates.

Machine learning with naturally labeled data for identifying abbreviation definitions, Yeganova L, Comeau DC, Wilbur WJ.

Abbreviations appear abundantly in biomedical text. Therefore, detecting abbreviations and identifying their definitions is important for accurate text analysis. Studies have estimated that tens of thousands of new abbreviations are added to Medline every year. Moreover, about 81% of abbreviations in Medline are ambiguous. In the paper, Yeganova et al. [3] use machine learning methods to address this problem. The purpose of the study is two-fold:

1) Create a machine learning approach that would achieve better accuracy than existing state-of the-art rule-based methods

2) Make use of naturally labelled data for training.

The key idea here is that training data is not labelled manually but is created automatically. Naturally occurring abbreviation-definition pairs in text are used to acquire positive training examples, while negative examples are generated randomly.

A structural SVM approach for reference parsing, Zhang, X, Zou J, Le DX, Thoma GR.

Automatic extraction of bibliographic data, such as article titles, author names, abstracts, and references are essential to citation databases such as MEDLINE. In their paper Zhang et al. [9] examine the task of parsing or identifying the components of bibliographic references. That information is essential for automatic indexing and in addition may provide valuable information for subsequent information extraction. For example, that knowledge may be used in identifying and extracting pairs of articles in a citing-cited relationship.

Due to strong regular internal structure, the problem is treated as a sequence labeling problem. A structural SVM is applied to solve the problem and the performance is compared to an earlier implementation of Conditional Random Fields approach. The authors provide a comparison between these machine learning algorithms, contrasting their performance results, pointing out their strengths and differences.

Acknowledgements

Funding: This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

This article has been published as part of *BMC Bioinformatics* Volume 12 Supplement 3, 2011: Machine Learning for Biomedical Literature Analysis and Text Retrieval. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/12?issue=S3.

Authors' contributions

LY and RID are the special session co-chairs at ICMLA 2010 and contributed equally to the overall organization, reviewing and editing of this supplement on "Machine Learning for Biomedical Literature Analysis and Text Retrieval".

Competing interests

The authors declare that they have no competing interests.

Published: 9 June 2011

References

1. Lu Z: **PubMed and beyond: a survey of web tools for searching biomedical literature.** *Database* Oxford; 2011.
2. Barrett N, Weber-Jahnke J: **Building a biomedical tokenizer using the token lattice design pattern and the adapted Viterbi algorithm.** *BMC Bioinformatics* 2011, **12(Suppl 3)**:S1.
3. Yeganova L, Comeau DC, Wilbur WJ: **Machine learning with naturally labeled data for identifying abbreviation definitions.** *BMC Bioinformatics* 2011, **12(Suppl 3)**:S6.
4. Islamaj Doğan R, Névéol A, Lu Z: **A context-blocks model for identifying clinical relationships in patient records.** *BMC Bioinformatics* 2011, **12(Suppl 3)**:S3.
5. Benton A, Hill S, Ungar L, Chung AP, Leonard C, Freeman C, Holmes JH: **A system for de-identifying medical message board text.** *BMC Bioinformatics* 2011, **12(Suppl 3)**:S2.
6. Aronson AR: **Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program.** *Proceedings of AMIA symp* 2001, 17-21.
7. Unified Medical Language System® (UMLS®):[http://www.nlm.nih.gov/research/umls/].
8. Jimeno-Yepes A, McInnes BT, Aronson AR: **Collocation analysis for UMLS knowledge-based word sense disambiguation.** *BMC Bioinformatics* 2011, **12(Suppl 3)**:S4, In this supplement.
9. Zhang X, Zou J, Le DX, Thoma GR: **A structural SVM approach for reference parsing.** *BMC Bioinformatics* 2011, **12(Suppl 3)**:S7.
10. Wilbur WJ, Kim W: **Improving a gold standard: treating human relevance judgments of MEDLINE document pairs.** *BMC Bioinformatics* 2011, **12(Suppl 3)**:S5.
11. Tanabe L, Wilbur WJ: **A Priority Model for Named Entities.** *BioNLP '06 Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis* 2006.

doi:10.1186/1471-2105-12-S3-I1
Cite this article as: Islamaj Doğan and Yeganova: **Topics in machine learning for biomedical literature analysis and text retrieval.** *BMC Bioinformatics* 2011 **12**(Suppl 3):I1.