Automatically Identifying Key Sentences in Biomedical Abstracts Using Semi-Supervised Learning

Bachelor thesis exposé

Martin Wackerbauer
Supervisor: Jurica Ševa

30th September 2017

Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät II
Institut für Informatik

Contents

Motivation

The ever-growing amount and availability of biomedical literature is a valuable source of information for clinical decisions. At the same time, it makes finding the right piece of information a cumbersome task: Clinicians are said to often read only a few sentences of an abstract to decide whether it is relevant to their work [MS03]. To support evidence-based medicine, text mining techniques are increasingly employed to find, condense and analyse relevant information [KMCY11].

As part of the "comPREhensive Data Integration for Cancer Treatment" (PREDICT) project [L+17], a software has been developed that finds clinical articles relevant to a patient's mutation profile, making the retrieval of such documents easier and more accurate than traditional keyword search. However, judging the returned documents' relevance to a specific clinical situation can still be time-consuming.

This work aims at automatically identifying key sentences in the abstracts of oncological articles. This will help clinicians quickly find the most relevant information and make informed decisions.

Problem Statement

The goal of this work is to develop a binary classifier that finds key sentences in oncological abstracts, i.e. the sentences most likely to provide a summary of clinical research. Given a corpus of labelled examples of positive and negative sentences, this can be implemented using standard supervised machine learning techniques to train the distinction between the two classes. However, a gold standard corpus is not available for the problem at hand, so we have to employ alternative approaches to generate training data. In doing so, we will use three sources of data:

- the CIViC Clinical Evidence Summaries database, a collection of summaries of PubMed articles on gene variant interpretations [GSK+17];
- the (unlabelled) abstracts corresponding to summaries in CIViC;
- the PIBOSO corpus used in [KMCY11], consisting of MEDLINE articles with sentences labelled according to their relevance to the standard criteria of evidence-based medicine.

The CIViC Clinical Evidence Summaries provide an initial set of reliable positive examples. These summaries, however, lack some of the most discriminant features for sentence classification, such as sentence location and context [TM02, MS03], which in turn are present in the corresponding abstracts' unlabelled sentences. The sentences labelled as *Other* in the PIBOSO corpus, i.e. not relevant to evidence-based medicine, can be used as examples of our negative class.

Our work will consist of two steps:

1. Using semi-supervised learning, we will generate a silver standard corpus based on CIViC summaries and unlabelled sentences in the corresponding set of PubMed abstracts. Since this is an active field of research with inherent uncertainties, we will explore different approaches to learning and validation.

2. Based on the constructed silver standard, we will use an existing pipeline to train a supervised machine learning algorithm for classifying new sentences. To optimise the classifier's reliability, we will use cross-validation and evaluate the effects of varying combinations of features.

To provide optimal benefit to clinicians, the results will be included in an extension to a web application that enables users to search for documents relevant to a given mutation. The new classifier will be used to highlight sentences in the returned abstracts according to their relevance score.

Data

CIViC is an open resource for Clinical Interpretation of Variants in Cancer [GSK+17]. The CIViC Clinical Evidence Summaries database consists of some 2000 evidence statements summarising findings about gene variants, such as their association with diseases and the outcome of drug response trials. Each entry contains a citation and the PubMed ID of the article the information is taken from. Additional information includes the names of the respective genes, variants, drugs, diseases, and a variant summary, which could be useful in the corpus generation step. These evidence statements are prototypes of high-quality information in condensed form and will make up our initial set of positive sentences. The corresponding abstracts can be retrieved as plain text via their PubMed IDs and will serve as unlabelled data for semi-supervised learning.

Additionally, the PIBOSO corpus used in [KMCY11] was kindly provided by the authors. It consists of 1000 biomedical abstracts with each sentence manually annotated according to the categories *Population, Intervention, Background, Outcome, Study Design*, and *Other* (i.e. not relevant to any of the standard criteria of evidence-based medicine). While there may not be a complete and unambiguous mapping from this classification to our own, sentences that are labelled only as *Other* can be regarded as not summarising clinical evidence. Thus, we can use them as auxiliary negative examples in constructing and/or validating a silver standard. Unfortunately, we cannot confidently use positive sentences from the remaining classes, since the classified abstracts are from various areas of medicine, while we focus on oncology.

Related Work

Sentence classification has been used in a wide range of fields, among them sentiment analysis [YH03, GBH09], rhetorical annotation, and automated summaries [KPC95, TM02]. The sparsity of labelled data and the cost of manually labelling text, with unlabelled text being abundant, has sparked interest in alternative approaches to training and corpus generation in text mining.

Sentence Classification

A special case of text categorisation, sentence classification shares some properties with document classification. For instance, sentences are usually represented as high-dimensional bag-of-word (BOW) or n-gram vectors, with classes often being linearly separable, and the same types of classifiers yield good results, such as Support Vector Machines (SVMs) [Joa98] and Convolutional Neural Networks [ZRW16].

However, there are some aspects specific to sentence classification, particularly regarding feature engineering. Whereas in text categorisation, techniques such as stop word removal, stemming or lemmatisation, and weighting terms by $tf * idf$ typically improve precision and recall, they may be ineffective or even counter-productive in sentence classification. For instance, common stop words like “but”, “was”, and “has” are often among the top features and verb tense, eliminated by stemming, can play an important role

in the rhetorical status of a sentence [AY09, KMA06]. Various feature selection heuristics exist for reducing the feature space's dimensionality while retaining the most useful information.

[TM02] have proposed a set of features for identifying rhetorical roles of sentences in summarising scientific full-text articles, among them sentence length and location, the presence of citations and of words included in headlines, properties of previous sentences, and formulaic expressions, with sentence position found to be the most discriminant feature.

In the biomedical domain, [MS03] have classified sentences in abstracts as belonging to the categories *Introduction, Methods, Results*, or *Discussion* (IMRaD). Rather than annotation by hand, the section headlines of explicitly structured abstracts were used as soft labels for training data. The authors used a simple bag-of-words (BOW) representation, but report that adding sentence location as a feature greatly improved classification. Among the classifiers tested, SVMs with linear kernel performed best.

[KMCY11] have also classified biomedical abstracts, trained on the hand-annotated PIBOSO corpus. They used Conditional Random Fields for sequential classification and tested an extensive set of features, including bigrams, part-of-speech (POS) tags, UMLS Concept Unique Identifiers and synonym expansion, and the predicted labels of previous sentences. In their experiments, the use of bigrams and semantic information did not pay off – probably due to data sparsity –, while positional information, preceding sentences, and section headings were beneficial to classification.

Semi-Supervised Learning

Semi-supervised learning relies on only a small set of labelled data, exploiting the information in unlabelled examples. Research indicates that semi-supervised learning has the potential to match the performance of supervised learning while requiring considerably less labelled data.

In recent years, Reinforcement Learning (RL) [WvHPS11] and Imitation Learning [HIE12] have been applied to document classification. [FYF+16] have used Inverse Reinforcement Learning to infer rewards for unlabelled data from the trajectories of an agent trained in a supervised setting to generalise Reinforcement Learning to the semi-supervised case.

The semi-supervised techniques we are going to focus on usually rely on

supervised classifiers trained on labelled data and use unlabelled examples to improve their decision boundary.

Self-Training

Self-Training is a general wrapper method using standard supervised algorithms at its core. It can be described as follows:

1. train initial classifier on labelled data
2. predict labels for remaining unlabelled data
3. move samples with most confidently predicted labels from unlabelled set to training data
4. retrain classifier on new training set
5. repeat from (2) until terminating condition is met.

Naïve Bayes is a popular classifier for Self-Training because the probabilities it produces provide a confidence ranking, but other algorithms may be used [WSLZ08].

In the field of sentiment classification, [LDGY13] have proposed Reserved Self-Training as a method for tackling unbalanced distributions where there are more negative than positive examples. Here, a random reserved sample of the labelled set is inserted into the unlabelled data. For retraining, the least confidently predicted examples from the reserved set are added to the training data along with the most confidently predicted unlabelled examples, improving discrimination between the classes. The authors used an SVM classifier, with confidence measured in distance between a sample and the dividing hyperplane.

Learning From Positive and Unlabelled Examples

Learning from positive and unlabelled examples (PU Learning) can be seen as a special case of semi-supervised learning where only positive labelled data is initially available.

Conceptually simple approaches include one-class SVMs, which approximate the support of the positive class and treat negative examples as outliers, as

well as ranking methods, which rank unlabelled examples by their decreasing similarity to the mean positive example, as described in [MV14].

Another class of PU learning algorithms are two-step heuristics following the general approach:

1. determine a set of reliable negative examples from the unlabelled data
2. iteratively produce a sequence of classifiers and choose the best one by some heuristic.

The reliable negative set can be obtained for instance by making use of Naïve Bayes' ranking property and using a confidence threshold. This threshold may be predefined, or can be determined by inserting positive "spy documents" into the unlabelled data and observing their score. Step (2) can be done by using the expectation maximisation (EM) algorithm, which is iteratively retrained using probabilistic labels, similar to Self-Training. Alternatives have been proposed using different classifiers and heuristics, from iterating EM until convergence to running an SVM only once [LLYL02].

A different approach is the so-called Biased-SVM [LDL$^+$03], which views the unlabelled data as a noisy set of negative examples, contaminated with positive examples. Thus, discriminating between positive and unlabelled examples with a bias towards the positive class can be considered an approximation of the original problem. The proposed classifier is a soft-margin SVM that uses different weights for the two classes in order to favour "false" positives over false negatives, i.e. to expand the area of the positive class. Biased-SVM has been reported to generally perform better than two-step techniques, ranked methods, and one-class SVM, and is considered state-of the art in [MV14].

Methods

Preprocessing

To make plain text data suitable for machine learning, it must be transformed into feature vectors. We will evaluate varying sets of features, such as BOW-vectors, n-grams, part-of-speech-tags, as well as sentence position and context as additional features where available.

Biomedical terms should be mapped to their semantic classes to account for data sparsity and avoid overfitting to particular genes or diseases, as well as to alleviate differences between the oncological domain of CIViC and the more general biomedical domain of negative sentences in PIBOSO. Other terms such as numbers should be replaced by class names, as well, leading to a combination of words and concepts.

Much of this is available in tools such as the Natural Language Toolkit [BKL09]; the UMLS Metathesaurus and MetaMap [Aro01] provide an ontology and advanced biomedical natural language processing useful for identifying named entities and underlying concepts.

Generating a Silver Standard

For lack of a gold standard corpus of labelled positive and negative examples suitable to the task at hand, we will generate a silver standard. As a starting point, CIViC Clinical Evidence Summaries represent our labelled positive examples. The PIBOSO corpus serves as an additional resource for validation and possibly training, with its *Other* class serving as reliable negative examples. The PubMed abstracts referenced in CIViC are the unlabelled set we will assign labels to via transductive semi-supervised learning.

Since the corpus' reliability is of practical relevance, we will explore the suitability of different approaches to semi-supervised learning and iteratively optimise our use of the most promising candidate. A combination of CIViC and PIBOSO can be used as training data for learning from positive as well as negative examples, such as Self-Training. PU Learning requires only the CIViC summaries and corresponding unlabelled abstracts for training; in this case, the negative sentences are needed only for validating classifiers.

Building a Classifier for Unseen Documents

Based on the silver standard corpus, we will train a supervised classifier using an existing pipeline. To make use of valuable properties present in PubMed abstracts' sentences but missing in CIViC summaries, such as context and sentence position, we can infer missing features that were unavailable in the previous step by imputing the mean values in the corpus. To maximise precision and recall of the classifier, we will use cross-validation and parameter optimisation and study the benefits of varying sets of features.

Schedule

The timeframe for a bachelor's thesis is 4 months (16 weeks).

Weeks	Task
1-3	building a basic preprocessing pipeline
4-6	evaluation of semi-supervised learning techniques and choosing the most promising approach
7-9	optimisation and validation of semi-supervised preprocessing and classification pipeline, generating a silver standard corpus
10-11	supervised training on silver standard using existing pipeline
12-13	improving the classifier and choosing optimal feature set
14-16	incorporating results into web application, tweaks, written thesis

References

[Aro01] Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17, 2001.

[AY09] Shashank Agarwal and Hong Yu. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, 25(23):3174–3180, 2009.

[BKL09] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly, 2009.

[FYF+16] Chelsea Finn, Tianhe Yu, Justin Fu, Pieter Abbeel, and Sergey Levine. Generalizing skills with semi-supervised reinforcement learning. *CoRR*, 2016.

[GBH09] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(2009):12, 2009.

[GSK+17] Malachi Griffith, Nicholas C Spies, Kilannin Krysiak, Joshua F McMichael, Adam C Coffman, Arpad M Danos, Benjamin J Ainscough, Cody A Ramirez, Damian T Rieke, Lynzey Kujan, et al. Civic is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature genetics*, 49(2):170, 2017.

[HIE12] He He, Hal Daumé III, and Jason Eisner. Imitation learning by coaching. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25, Lake Tahoe, Nevada, United States*, pages 3158–3166, 2012.

[Joa98] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In Claire Nedellec and Céline Rouveirol, editors, *Machine Learning: ECML-98, Chemnitz, Germany*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer, 1998.

[KMA06] Anthony Khoo, Yuval Marom, and David Albrecht. Experiments with sentence classification. In *Proceedings of the 2006 Australasian language technology workshop*, pages 18–25, 2006.

[KMCY11] Su Kim, David Martínez, Lawrence Cavedon, and Lars Yencken. Automatic classification of sentences to support evidence based medicine. *BMC Bioinformatics*, 12(S-2):S5, 2011.

[KPC95] Julian Kupiec, Jan O. Pedersen, and Francine Chen. A trainable document summarizer. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *SIGIR'95, Seattle, Washington, USA*, pages 68–73. ACM Press, 1995.

[L+17] Ulf Leser et al. Predict | comprehensive data integration for cancer treatment. `http://predict.informatik.hu-berlin.de`, 2017. Accessed: 2017-09-01.

[LDGY13] Zhiguang Liu, Xishuang Dong, Yi Guan, and Jinfeng Yang. Reserved self-training: A semi-supervised sentiment classification method for chinese microblogs. In *Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan*, pages 455–462. Asian Federation of Natural Language Processing / ACL, 2013.

[LDL+03] Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S. Yu. Building text classifiers using positive and unlabeled examples. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), Melbourne, Florida, USA*, pages 179–188. IEEE Computer Society, 2003.

[LLYL02] Bing Liu, Wee Sun Lee, Philip S. Yu, and Xiaoli Li. Partially supervised classification of text documents. In Claude Sammut and Achim G. Hoffmann, editors, *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), Sydney, Australia*, pages 387–394. Morgan Kaufmann, 2002.

[MS03] Larry McKnight and Padmini Srinivasan. Categorization of sentence types in medical abstracts. In *American Medical Informatics Association Annual Symposium, Washington, DC, USA*. AMIA, 2003.

[MV14] Fantine Mordelet and Jean-Philippe Vert. A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209, 2014.

[TM02] Simone Teufel and Marc Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

[WSLZ08] Bin Wang, Bruce Spencer, Charles X. Ling, and Harry Zhang. Semi-supervised self-training for sentence subjectivity classification. In Sabine Bergler, editor, *Advances in Artificial Intelligence , 21st Conference of the Canadian Society for Computational Studies of Intelligence Windsor, Canada*, pages 344–355. Springer, 2008.

[WvHPS11] Marco A. Wiering, Hado van Hasselt, Auke-Dirk Pietersma, and Lambert Schomaker. Reinforcement learning algorithms for solving classification problems. In *2011 IEEE Symposium on Adaptive Dynamic Programming And Reinforcement Learning, Paris, France*, pages 91–96. IEEE, 2011.

[YH03] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136, 2003.

[ZRW16] Ye Zhang, Stephen Roller, and Byron C. Wallace. MGNC-CNN: A simple approach to exploiting multiple word embeddings for sentence classification. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, San Diego California, USA*, pages 1522–1527. The Association for Computational Linguistics, 2016.