Journal of Biomedical Informatics 62 (2016) 59–65

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Topic detection using paragraph vectors to support active learning in systematic reviews

Kazuma Hashimoto [a,1], Georgios Kontonatsios [b,1], Makoto Miwa [c], Sophia Ananiadou [b,*]

[a] Graduate School of Engineering, University of Tokyo, Tokyo, Japan
[b] School of Computer Science, National Centre for Text Mining, University of Manchester, Manchester, United Kingdom
[c] Department of Advanced Science and Technology, Toyota Technological Institute, Nagoya, Japan

ARTICLE INFO

Article history:
Received 26 January 2016
Revised 4 April 2016
Accepted 5 June 2016
Available online 10 June 2016

Keywords:
Systematic reviews
Citation screening
Topic modelling
Paragraph vectors
Document embeddings
Active learning

ABSTRACT

Systematic reviews require expert reviewers to manually screen thousands of citations in order to identify all relevant articles to the review. Active learning text classification is a supervised machine learning approach that has been shown to significantly reduce the manual annotation workload by semi-automating the citation screening process of systematic reviews. In this paper, we present a new topic detection method that induces an informative representation of studies, to improve the performance of the underlying active learner. Our proposed topic detection method uses a neural network-based vector space model to capture semantic similarities between documents. We firstly represent documents within the vector space, and cluster the documents into a predefined number of clusters. The centroids of the clusters are treated as latent topics. We then represent each document as a mixture of latent topics. For evaluation purposes, we employ the active learning strategy using both our novel topic detection method and a baseline topic model (i.e., Latent Dirichlet Allocation). Results obtained demonstrate that our method is able to achieve a high sensitivity of eligible studies and a significantly reduced manual annotation cost when compared to the baseline method. This observation is consistent across two clinical and three public health reviews. The tool introduced in this work is available from https://nactem.ac.uk/pvtopic/.

1. Introduction

Systematic reviews involve searching, screening and synthesising research evidence from multiple sources, in order to inform policy studies and guideline development [1]. In evidence-based medicine, systematic reviews are vital in guiding and informing clinical decisions, and in developing clinical and public health guidance [2]. In carrying out systematic reviews, it is critical to minimise potential bias by identifying all studies relevant to the review. This requires reviewers to exhaustively and systematically screen articles for pertinent research evidence, which can be extremely time-consuming and resource intensive [3].

To reduce the time and cost needed to complete the screening phase of a systematic review, researchers have explored the use of active learning text classification to semi-automatically exclude irrelevant studies while keeping a high proportion of eligible studies (i.e., sensitivity of at least 95%) in the final review [4–6]. Active learning text classification is an iterative process that incrementally learns to discriminate eligible from ineligible studies. The process starts with a small seed of manually labelled citations that is used to train an initial text classification model. The active learner will then iterate through several learning cycles to optimise its prediction accuracy. At each learning cycle, the active learner automatically classifies the remaining unlabelled citations. A sample of the automatically labelled citations is validated by an expert reviewer. Finally, the validated sample is used to update (re-train) the classification model. The process terminates when a convergence criterion is satisfied (e.g., 95% of eligible studies is identified by the active learner).

Key to the success of the active learning approach is the feature extraction method that encodes documents into a vector representation that is subsequently used to train the text classification model. Wallace et al. [5] proposed a multi-view active learning approach that represents documents using different feature spaces, e.g., words that appear in the title and in the abstract, keywords and MeSH terms. Each distinct feature space is used to train a sub-classifier, e.g. Support Vector Machines (SVM). Multiple sub-classifiers are then combined into an ensemble classifier using a heuristic (e.g., majority votes). With regard to the active learning selection criterion (i.e., a function that determines the next sample

* Corresponding author.
E-mail address: sophia.ananiadou@manchester.ac.uk (S. Ananiadou).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jbi.2016.06.001
1532-0464/

of instances to be validated by the reviewer), the authors employed uncertainty sampling. The uncertainty selection criterion selects those instances for which the classifier is least certain of their classification label. To enhance the performance of the active learner, they introduced an aggressive undersampling technique that removes ineligible studies from the training set which convey little information. The aggressive undersampling technique aims at reducing the negative effect of class imbalance that occurs in systematic reviews, i.e., a high percentage of ineligible studies tends to overwhelm the training process. For experimentation, they applied the proposed method to three clinical systematic review datasets. They showed that the uncertainty-based active learner with aggressive undersampling is able to decrease the human-workload involved in the screening phase of a systematic review by 40–50%.

Whilst good results are obtained in the clinical domain, Miwa et al. [4] demonstrated that the active learning approach yields a significantly lower performance when applied to public health reviews. The authors argued that the identification of relevant studies is more challenging in this domain compared to others, e.g., clinical documents. This can be attributed to the fact that the public health literature extends across a wide range of disciplines covering diverse topics (e.g., social science, occupational health, education, etc.) [7]. To alleviate problems introduced by challenging public health articles, the authors proposed to learn a topic-based representation of studies by employing the widely used Latent Dirichlet Allocation (LDA) [8], a probabilistic and fully generative topic model. They further investigated the use of a certainty-based selection criterion that determines a validation sample consisting of instances with a high probability of being relevant to the review (as opposed to the previously introduced uncertainty sampling [5] that selects instances with low classification probability). Experimental results determined that topic-based features can improve the performance of the active learner. Moreover, the certainty-based active learner that uses topic features induced by LDA exceeded state-of-the-art performance and outperformed the uncertainty-based active learner [5].

Topic models are machine learning methods that aim to uncover thematic structures hidden in text. One of the earliest topic modelling methods is the probabilistic Latent Semantic Indexing (PLSI) [9]. PLSI associates a set of latent topics Z with a set of documents D and a set of words W (D, W are observed variables). The goal is to determine those latent topics that best describe the observed data. In PLSI the probability distribution of latent topics is estimated independently for each document. In practice, this means that the complexity of the model (i.e., number of parameters to be computed) grows linearly with the size of the collection. A further disadvantage of PLSI is the inability of the underlying model to generalise on new, unseen documents (i.e. the model is not fully generative). Extending upon of PLSI, LDA assumes that topic distributions are drawn from the same prior distribution which allows the model to scale up to large datasets and better generalise to unseen documents.

In this article, we present a novel topic detection model to accelerate the performance of the active learning text classification model used for citation screening. Our topic detection method can be used as an alternative approach to the LDA topic model to generate a topic-based feature representation of documents. The proposed method uses a neural network model, i.e., paragraph vectors [10], to learn a low dimensional, but informative, vector representation of both words and documents, which allows detection of semantic similarities between them. Previous work has demonstrated that paragraph vector models can accurately compute semantic relatedness between textual units of varying lengths, i.e., words, phrases [11] and longer sequences, e.g., sentences, paragraphs and documents [10]. While the standard bag-of-words approach (i.e., a document is represented as a vector of the words that it contains) has been frequently employed in various natural language processing tasks (e.g., text classification, sentiment analysis), paragraph vectors, which take into account factors such as word ordering within text, have been shown to yield superior performance [10].

To our knowledge, our work is the first that utilises the vector representations of documents produced by the paragraph vector model for topic detection. We hypothesise that documents lying close to each other in the vector space form topically coherent clusters. Based on this, our approach clusters the paragraph vector representations of documents by applying the k-means clustering algorithm and treats the centroids of the clusters as representatives of latent topics, assuming that each cluster corresponds to a latent topic inherent in the texts. After detecting latent topics in a collection of documents, we represent each document as a k-dimensional feature vector by calculating the distance of the document to the k cluster centroids. Additionally, our topic detection model computes the conditional probability that a word is generated by a given topic and thus readily determines a set of representative keywords to describe each topic. The topic-based representation of documents is used to train an active learning text classification model to more efficiently identify eligible studies for inclusion in a review. The contributions that we make in this paper can be summarised in the following points:

1. We propose a novel topic detection method that builds upon the paragraph vector model. We introduce various adaptations to the paragraph vector method that enable the underlying model to discover latent topics in a collection of documents and summarise the content of each topic by meaningful and comprehensive text labels.
2. We incorporate the new topic detection method with an active learning strategy to support the screening process of systematic reviews.
3. We conduct experiments, demonstrating that our topic detection method outperforms an existing topic modelling approach when applied to semi-automatic citation screening of clinical and public health reviews.

2. Methods

In this section, we detail our proposed topic detection method. We then provide an overview of the active learning process used in our experiments and discuss the evaluation protocol that we follow to asses the paragraph vector-based topic detection method.

2.1. A paragraph vector-based topic detection method

2.1.1. Word vectors

Several approaches on representing the meaning of words using mathematical expressions such as vectors and matrices have been proposed, with neural network models recently gaining much attention [11–13]. Neural network models usually fully parameterise the word vectors; in other words, each word w has n parameters in its word vector: $v(w) = (x_{w1}, x_{w2}, \ldots, x_{wn})$. The parameters are used to estimate the conditional probability that a target words will appear, given its context words. The parameters for each word are initialised with random values, and then adjusted in the learning process whose objective is to maximise the conditional probability:

$$p(w_t | w_{t-N}, w_{t-N+1}, \ldots, w_{t-1}) \quad (1)$$

where w_t is the target word and $w_{t-N}, w_{t-N+1}, \ldots, w_{t-1}$ are N context words that occur before w_t. During the learning process, the parameterised vectors of the context words are used and updated, and the