Evaluating topic model interpretability from a primary care physician perspective

Corey W. Arnold*, Andrea Oh, Shawn Chen, William Speier

Medical Imaging and Informatics Group, Department of Radiological Sciences, University of California, Los Angeles, United States

ARTICLE INFO

ABSTRACT

Background and objective: Probabilistic topic models provide an unsupervised method for analyzing unstructured text. These models discover semantically coherent combinations of words (topics) that could be integrated in a clinical automatic summarization system for primary care physicians performing chart review. However, the human interpretability of topics discovered from clinical reports is unknown. Our objective is to assess the coherence of topics and their ability to represent the contents of clinical reports from a primary care physician's point of view.

Methods: Three latent Dirichlet allocation models (50 topics, 100 topics, and 150 topics) were fit to a large collection of clinical reports. Topics were manually evaluated by primary care physicians and graduate students. Wilcoxon Signed-Rank Tests for Paired Samples were used to evaluate differences between different topic models, while differences in performance between students and primary care physicians (PCPs) were tested using Mann–Whitney *U* tests for each of the tasks.

Results: While the 150-topic model produced the best log likelihood, participants were most accurate at identifying words that did not belong in topics learned by the 100-topic model, suggesting that 100 topics provides better relative granularity of discovered semantic themes for the data set used in this study. Models were comparable in their ability to represent the contents of documents. Primary care physicians significantly outperformed students in both tasks.

Conclusion: This work establishes a baseline of interpretability for topic models trained with clinical reports, and provides insights on the appropriateness of using topic models for informatics applications. Our results indicate that PCPs find discovered topics more coherent and representative of clinical reports relative to students, warranting further research into their use for automatic summarization.

© 2015 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

The primary care physician's (PCP) role is to deliver comprehensive care to their patients. Irrespective of the complexity of a patient's medical history, the PCP is responsible for organizing and understanding relevant problems to make informed decisions regarding care. Unfortunately, PCPs have high time demands, with a large portion of time involved in indirect patient care (reading, writing, and searching for data in

---

* *Corresponding author at*: 924 Westwood Blvd Ste 420, Los Angeles, CA 90024, United States. Tel.: +1 310 794 3538; fax: +1 310 794 3546.
  E-mail address: cwarnold@ucla.edu (C.W. Arnold).

support of patient care) [1]. While the use of automated tools and overviews/summaries for patient records have been studied to facilitate this time consuming process, efforts have been limited to a narrow range of tasks and basic, superficial temporal representations [2–6]. As physicians continue to struggle with how much control they have over their time to deliver an increasing number of services and patient-centered care in managerially driven organizations, they would benefit from utilities that expedite the medical chart review process by providing meaningful automated summarization that assists in answering clinical questions [7]. The development of a model that captures the expression of key concepts could help alleviate some of the time burden felt by PCPs.

Automatic summarization of clinical documents is an active area of research, both in general [8,9] and specifically in the clinical domain [10]. A key component of developing an automatic summarization system is finding concept similarity, which represents abstract connection between different words beyond their usage and meaning [10]. In small, well-defined domains, it has been shown that ontology-based methods work well [11,12], but it remains an open problem in broader domains and in general has not been translated to most summarization systems [10]. Topic modeling is a method designed for identifying such abstract connections, so it could potentially be leveraged to achieve concept similarity for summarization systems in the clinical domain.

Probabilistic topic models for language have been widely explored in the literature as unsupervised, generative methods for quantitatively characterizing unstructured free-text with semantic topics. These models have been largely discussed for general corpora (e.g., newspaper articles), and have been developed for many uses, including word-sense disambiguation [13], topic correlation [14], learning information hierarchies [15], and tracking themes over time [16,17]. In the biomedical domain, work has investigated the use of topic models to evaluate the impact of copy and pasted text on topic learning [18], better understanding and predicting Medical Subject Headings (MeSH) applied to PubMed articles [19], and exploring the correlation between Federal Drug Administration (FDA) research priorities and topics in research articles funded under those priorities [20]. Recently, topic models have been employed in the clinical domain in problems such as cased-based retrieval [21]; characterizing clinical concepts over time [22]; and predicting patient satisfaction [23], depression [24], infection [25], and mortality [26]. Additional work has been performed in using topic modeling methods to search for relationships between themes discovered in clinical notes and underlying patient genetics [27].

Exposing topics directly to PCPs through an integrated visualization is a possible mechanism for automatic summarization and information filtering of clinical records [28]. However, such a system would require that topics are human-interpretable and accurately reflect the contents of the medical record. While there has been work to evaluate the interpretability of topic models for general text collections [29,30], no work has investigated the ability of a topic model to extract human-interpretable topics from clinical free-text. Clinical documents pose additional challenges in that they contain specialized information that requires significant training and experience to understand. As a result,

using lay people as evaluators is probably insufficient for a clinical topic model as they would underperform due to a lack of domain knowledge rather than a lack of topic coherence.

In this paper, we present such an evaluation and compare the results of a topic model at several levels of granularity as interpreted by PCPs and lay people. While previous studies have had physicians evaluate topics from clinical text [24], to the best of our knowledge, no work has sought to compare topic interpretability between target users (PCPs) and baseline laypersons as method for evaluating the ability of a topic model to capture specialized themes. Our goal is to establish that a basic topic model is capable of discovering coherent topics that are representative of clinical reports.

2.   Background

Seminal work in exploring latent semantics in free-text includes latent semantic indexing (LSI) [31], which applies singular value decomposition (SVD) to a weighted term-document matrix to arrive at a lower-rank factorization that can be used to compare the similarity of terms or documents. Through the contextual co-occurrence patterns of words in the matrix, the technique can overcome the problems of synonymy and polysemy. Probabilistic LSI (PLSI) [32] models the joint distribution of documents and words using a set of latent classes. Each document is represented as a mixture model of latent classes ("topics") that are defined as multinomial distributions over words. Thus, generating a word requires selecting a latent class based on its proportion in the document and then sampling a word based on that latent class' word distribution. The model is fit using the expectation maximization (EM) algorithm [33].

Latent Dirichlet allocation (LDA) [34] is a bag-of-words model that is similar to PLSI in that documents are mixtures of word-generating topics. However, LDA goes a step further and proposes a generative model for document-topic mixtures using a Dirichlet prior on a document's topic distribution. LDA assumes topics exist in a Dirichlet-distributed latent space, from which document multinomial topic mixtures are drawn. A topic may then be sampled from the topic multinomial, which indexes individual topics from which words are drawn to generate documents. The inclusion of a Dirichlet prior has the benefit of mitigating overfitting, which is a limitation of PLSI [18].

2.1.   Topic evaluation

LDA models are typically evaluated by computing the likelihood that a held-out document set was generated by a model fit to training documents [35]. Such likelihood metrics are objective and generalize well to different model configurations and data collections. In addition, they provide performance feedback during parameter fitting. Indirect evaluation of topic models may be performed in combination with a classifier, such as a support vector machine (SVM) model, trained on topic model generated features for a particular task.

While the above evaluations inform how well a topic model fits to data (under model assumptions) and the utility