Title
Evaluating topic model interpretability from a primary care physician perspective

Authors
Arnold, Corey W
Oh, Andrea
Chen, Shawn
et al.

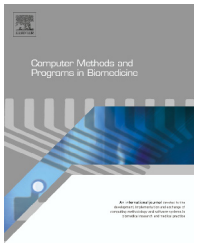
Publication Date
2016-02-01

DOI
10.1016/j.cmpb.2015.10.014

Peer reviewed

COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE XXX (2015) XXX–XXX

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



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

Article history:
Received 12 June 2015
Received in revised form
14 September 2015
Accepted 20 October 2015

Keywords:
Topic modeling
Primary care
Clinical reports

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.

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).

http://dx.doi.org/10.1016/j.cmpb.2015.10.014
0169-2607/

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

Please cite this article in press as: C.W. Arnold, et al., Evaluating topic model interpretability from a primary care physician perspective, Comput. Methods Programs Biomed. (2015), http://dx.doi.org/10.1016/j.cmpb.2015.10.014

of learned topics for subsequent tasks, they do not explicitly measure the human-interpretability of discovered topics. Judging the human interpretability of topics is more challenging, as each person may have different notions of how words interrelate and the meaning they convey. Such subjectivity makes the direct use of topics difficult, as it is uncertain if users will interpret them to mean the same thing when using a system. There has been some work to explore the quality of the semantic representations discovered by topic models [15]. Notably, [29] used Amazon's Mechanical Turk, an online service to scale task completion by human workers [36], to compare the interpretability over several models using the two tasks of word intrusion and topic intrusion. These tasks measure a topic model's ability to generate coherent topics and representative document topic mixtures.

3. Methods

3.1. Data collection

Our data collection consisted of medical reports for patients with brain cancer, lung cancer, or acute ischemic stroke collected by identifying patients from an IRB-approved disease-coded research database. In total, the data set consisted of 936 patients, with a total of 84,201 medical reports. As our use-case of interest for the topic models was to support an automatic summarization system for PCPs, we filtered this collection by report type to select those reports composed primarily of uncontrolled free-text that summarize a patient's episode of care. When evaluating a new patient, PCPs perform a process of information summarization through chart review using these types of reports, selectively drilling-down to more specific information when needing to address a specific clinical question. We therefore selected progress notes, consultation notes, history and physicals (H&Ps), discharge summaries, and operative reports/procedures/post-op notes from our total set of reports, resulting in 20,161 reports from 924 patients (12 patients did not have a report of interest) that we used to fit topic models. The median number of reports for a patient was 13. The minimum was one report and the maximum was 260 reports. These reports were preprocessed to remove punctuation, stop words, words that occurred in fewer than five documents, and words that occur in every document. Protected health information (PHI) and numbers were also removed automatically using regular expressions. The resulting dataset consisted of 17,993 unique tokens and 5,820,160 total tokens.

3.2. Topic models

In this work, we sought to establish a baseline level of topic interpretability using a general form of topic model, LDA. To compare the interpretability of LDA topics with differing degrees of granularity, three models were fit: one with 50 topics, one with 100 topics, and one with 150 topics. Models were fit in 4000 iterations of Gibbs sampling using MALLET software [37], which was configured to use hyperparameter optimization, a 200 iteration burn-in period, and a lag of 10.

Table 1 – Top five words from five topics for each model and the random intruder word used in the word intrusion task.

#	Top words	Intruder
50 topics		
24	intact, scan, mri, normal, ph	liver
26	radiation, treatment, therapy, oncology, dose	bid
33	allergies, disease, family, cancer, social	lobe
34	lung, lobe, upper, lymph, surgery	abdominal
39	prn, mca, bid, iv, daily	family
100 topics		
2	brain, ct, mca, infarct, mri	sodium
6	assistance, daily, rehabilitation, mobility, activities	night
8	trial, daily, bevacizumab, treatment, irinotecan	problems
23	count, blood, sodium, potassium, hemoglobin	prostate
61	femoral, lower, vascular, extremity, foot	recurrence
150 topics		
3	effusion, pleural, pericardial, daily, moderate	abdominal
11	operation, placed, surgeon, incision, using	clot
28	pain, emergency, room, episode, blood	cath
116	liver, hepatitis, colon, cirrhosis, post	commands
118	rate, exercise, stress, heart, normal	clear

3.3. Model likelihood

To compare goodness-of-fit across models, the document collection was randomly split, with 80% of all documents used for training and 20% held out for testing. After training, the held-out log-likelihood was then computed for each model using "left-to-right" sampling [35], which estimates $P(\mathbf{W}_{\text{test}}/\alpha,\beta)$. Resulting likelihoods were normalized by the number of tokens in their respective set. The split into train and test sets was used only for computing predictive model likelihood for comparison across models. The subsequent word and topic intrusion tasks were performed using models fitted to all documents.

3.4. Word intrusion task

To gauge the human interpretability of topics, we used the tasks of word intrusion and topic intrusion. In the word intrusion task subjects look at a list of six randomly ordered words and are asked to select the word that "does not belong." Five of the words are the most probable words given the topic, with the other word having a low probability in the topic. Low probability words are selected by randomly choosing from the least probable 20 words in a topic, with the constraint that the selected word must also be a top five word probability-wise in a different topic. Table 1 shows five example word intrusion questions from each topic model. The precision of a topic model is then the fraction of subjects who correctly identify the intruding word. For model m and topic t, let ω_t^m be the index of the intruding word and $i_{t,s}^m$ be the index of the intruder selected by subject s, with S representing the total number of subjects. Model precision (MP) may then be calculated as:

$$\text{MP}_t^m = \frac{1}{S}\sum_s 1(\omega_t^m = i_{t,s}^m)$$

Please cite this article in press as: C.W. Arnold, et al., Evaluating topic model interpretability from a primary care physician perspective, Comput. Methods Programs Biomed. (2015), http://dx.doi.org/10.1016/j.cmpb.2015.10.014

Table 2 – Topic intrusion task example. Participants are asked to pick the topic that does not belong in the accompanying report. In this example, topic 7 is the intruder.

Report (truncated)

The patient is a YY-year-old female seen for continuing care following aortic valve replacement. Today, she has had a low cardiac index although interestingly, she is not acidotic, is mentating well, and is making good urine without the use of diuretics. Thus, this appears to be acceptable for her. We have given the patient aggressive fluid resuscitation in order to help improve her left ventricular outflow tract obstruction. We have gingerly started a small dose of afterload reduction to see if this will help, although this may be problematic in the setting of LVOT obstruction. The patient will be observed carefully. We will likely be able to discontinue her Swan later on today or tomorrow. The patient has been extubated. We will obtain a swallowing...

Randomly ordered topics

93	78	7	23
pulmonary	clear	mca	count
status	rate	cx	blood
respiratory	extremities	artery	sodium
failure	abdomen	infarct	potassium
intubated	nontender	trach	hemoglobin

3.5. Topic intrusion task

In the topic intrusion task, subjects are shown four randomly ordered topics and a report. Three of the topics are those that best summarize the words contained in the displayed report, and one of the topics has a low probability in the report (selected similarly to the word intrusion task). Subjects are asked to select the topic that "does not belong" in the report. As seen in Table 2, for each topic, the five most probable words are displayed (i.e., there are four lists, each with five words). For model m, let $\hat{\theta}_d^m$ represent the point estimate of the topic proportions vector for document d, and let $\hat{\theta}_{d,s}^m$ be the intruding topic selected by subject s and $\hat{\theta}_{d,*}^m$ be the true intruding topic, with S representing the total number of subjects. The topic log odds (TLO) may then be calculated as:

$$\mathrm{TLO}_d^m = \frac{1}{S}\sum_s \log \hat{\theta}_{d,*}^m - \log \hat{\theta}_{d,s}^m$$

3.6. Topic evaluators

In order to detect if the topic model is learning specialized clinical themes, we sought to compare the performance of PCPs with baseline laypersons who had no formal training in medicine. Therefore, PCPs and informatics students were surveyed on the interpretability of topics from each of the three models using the word intrusion and topic intrusion tasks. Thus, 10 randomly selected topics from each model and five randomly selected reports were evaluated by each subject for the word intrusion task and the topic intrusion task, respectively. Subjects were recruited via email and evaluations were performed using a web interface, with each subject receiving identical instructions on how to complete the tasks.

Table 3 – Held-out log-likelihood scores for different LDA models fit to clinical text collection.

# Topics	Held-out log-likelihood
50	−7.1951
100	−7.0906
150	−7.0360

3.7. Statistical analysis

Model precision and topic intrusion scores were averaged over trials and subjects. Additionally, the time taken to select an answer was recorded for each subject and the median was found for each model size. For each of the tasks, performances between the models were compared using Wilcoxon Signed-Rank Tests for Paired Samples with alpha equal to 0.05. Differences in performance between students and PCPs were tested using Mann–Whitney U tests for each of the tasks.

3.8. Automated topic coherence

To investigate if student and PCP determinations of interpretability correspond with automatic measures of topic coherence, we calculated the pointwise mutual information (PMI) for each topic. PMI has been used previously in evaluating topic models [30,38], and measures the statistical independence of observing two words in close proximity within a text collection. Following previous work, we analyze the top 10 words within each topic, and determine their co-occurrence using a sliding window of 10 words within the corpus of clinical reports. PMI is then calculated over the 10-word windows as follows:

$$\mathrm{PMI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i) * p(w_j)}$$

Using the top 10 words from each topic results in (10/2) (i.e., 45) PMI scores per topic, which were then averaged and compared to average word intrusion results for PCPs and students using the Spearman correlation coefficient.

4. Results

Models required approximately 75–90 min to train on a desktop-grade computer. Table 3 details the results of the held-out log-likelihood analysis. Using log-likelihood, LDA is better able to fit to the data as the number of topics increases, a trend also observed in the literature [29]. Additionally, we noticed that topics, as expected, became more granular as the number of topics increased. For example, the 50 topic model has a single topic for lung disease (pulmonary, pneumonia, pleural, lung, effusion) while the 150 topic model has separate topics for pleural effusion (effusion, pleural, pericardial, daily, moderate) and pneumonia (pulmonary, lung, pneumonia, lobe, cough). The 150 topic model also has separate topics for pacemaker (pacemaker, lead, atrial, ventricular, pacing) and valve repair (valve, aortic, repair, post, endocarditis) while the 50 topic model combines the two (repair, pacemaker, valve, closure, post).

Please cite this article in press as: C.W. Arnold, et al., Evaluating topic model interpretability from a primary care physician perspective, Comput. Methods Programs Biomed. (2015), http://dx.doi.org/10.1016/j.cmpb.2015.10.014

Table 4 – Example PCP and student topic precision results with median PMI score for each model.

# Topics	PCP precision	Student precision	Median PMI	Topic	Top words	Intruding word
	0.94	0.81	0.31	18	coronary, disease, artery, pressure, cardiac	intact
50	0.58	0.27	0.88	34	lung, lobe, upper, lymph, surgery	abdominal
	0	0.1	−0.08	19	prn, mca, bid, iv, daily	course
	1	0.63	0.97	61	femoral, lower, vascular, extremity, foot	recurrence
100	0.65	0.45	−0.09	68	continue, plan, bid, bp, hr	scan
	0	0	−0.90	52	intact, scan, mri, ph, normal	large
	0.94	0.64	1.05	12	surgery, risks, discussed, surgical, benefits	cxr
150	0.53	0.45	0.25	28	pain, emergency, room, episode, blood	cath
	0.18	0	−0.10	120	respiratory, blood, normal, neurologic, care	mri

4.1. *Word intrusion task*

For the word intrusion task, there were 28 total respondents composed of 17 PCPs and 11 informatics students. Table 4 shows example topics and their precision results from each model for PCPs and students. Both PCPs and students performed best in terms of both time and precision on the 100 topic model. The median precision of PCPs using the 100 topic model was 70% and the median time to make a selection was 9.5 s. This performance was significantly better than the 50 and 150 topic models in terms of precision ($p = 0.001$ and $p = 0.01$, respectively), but the improvement was not statistically significant in terms of time ($p = 0.06$ and $p = 0.20$, respectively) (Fig. 1a, c).

Students took a median time of 11.5 s and achieved a median precision of 60% using the 100 topic model. This performance was significantly better than the 50 and 150 topic models in terms of time ($p = 0.045$ and $p = 0.045$, respectively) (Fig. 1b, d). In terms of precision, it was significantly better than the 50 topic model ($p = 0.025$), but not the 150 topic model ($p = 0.091$). PCPs performed significantly better than students in terms of both time ($p = 0.027$) and model precision ($p = 0.015$).

4.2. *Topic intrusion task*

For the topic intrusion task, there were 20 total respondents, with 9 PCPs and 11 informatics students. As the topic

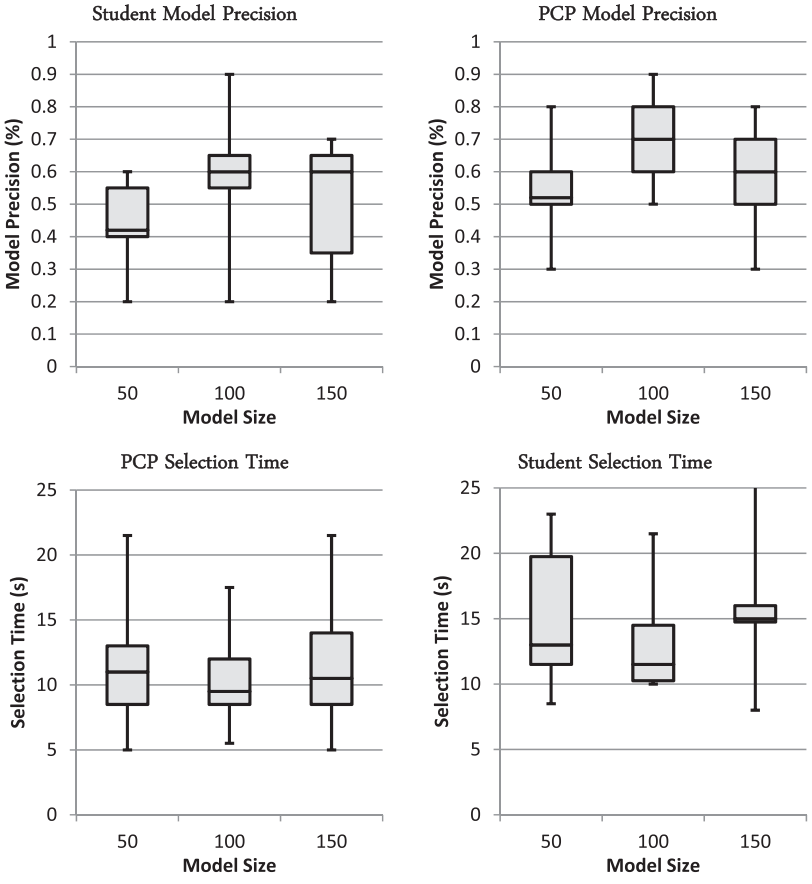


Fig. 1 – Time and precision box plots for word intrusion task. Median selection times and precisions were found for each topic model size for both primary care physicians (a, c) and informatics students (b, d) performing the word intrusion task.

Please cite this article in press as: C.W. Arnold, et al., Evaluating topic model interpretability from a primary care physician perspective, Comput. Methods Programs Biomed. (2015), http://dx.doi.org/10.1016/j.cmpb.2015.10.014

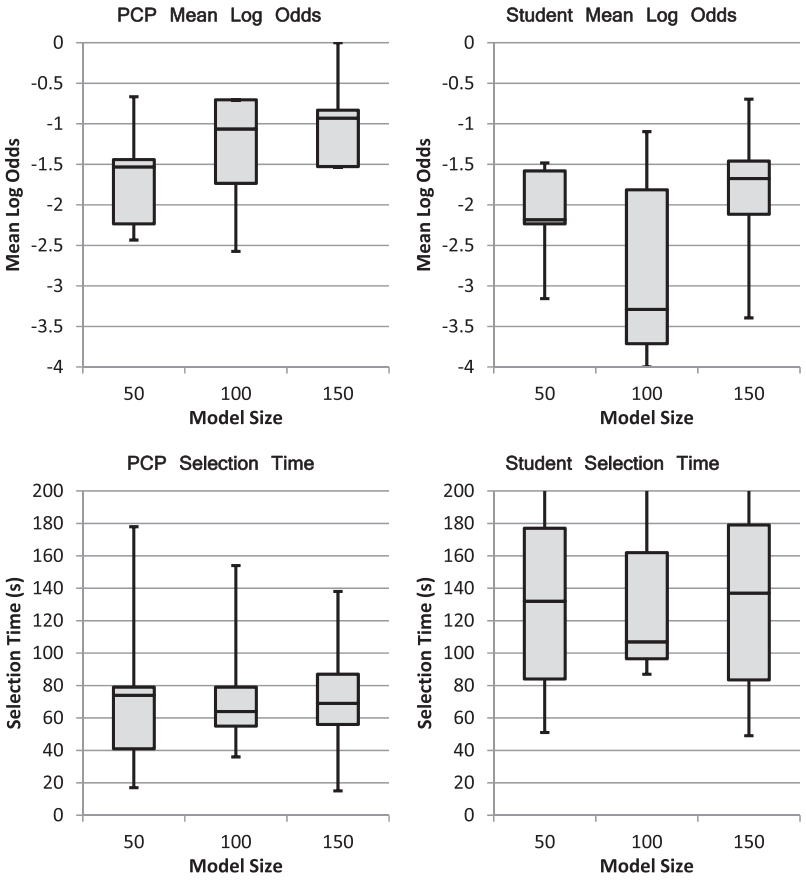


Fig. 2 – Time and log odds box plots for topic intrusion task. Median selection times and mean log odds were found for each topic model size for both primary care physicians (a, c) and informatics students (b, d) performing the topic intrusion task.

intrusion task is more time consuming than the word intrusion task, fewer PCPs participated. Both students and PCPs spent the least amount of time to make selections when using the 100 topic model. The median time required for PCPs to make selections using the 100 topic model was lower than that for the 50 topic model (74 vs. 64 s) and the log odds achieved was higher (−1.06 vs. −1.48), but neither were significantly different ($p = 0.47$ and $p = 0.21$, respectively) (Fig. 2a, c). The median time required for selections using the 150 topic model was slightly higher (69 s) and the median log odds was higher (−0.87), but these were also not significantly different from the PCP results for the 100 topic model.

Students also had the lowest median selection time using the 100 topic model (107 s), but it was not significantly different from the selection times using the 50 topic model (132 s, $p = 0.72$) nor the 150 topic model (137 s, $p = 0.92$) (Fig. 2d). However, students also had the lowest log odds for the 100 topic model (−3.29), which was significantly lower than that achieved using the 50 topic model (−2.23, $p = 0.04$) and the 150 topic model (−1.67, $p = 0.02$) (Fig. 2b). Again, doctors performed significantly better than students in terms of both time and log odds ($p = 0.0028$ and $p = 0.0039$, respectively).

Results of the topic intrusion task were highly dependent on the probability of the lowest probability relevant topic. In cases where the probability of this topic was low, subjects had a harder time distinguishing it from the intruding topic. Combining results across models, physicians had a median log odds of −0.18 for the eight documents with a third topic probability greater than 0.1 compared to a median log odds of −1.50 for the seven documents with a third topic probability less than 0.1. Students had a similar trend with a median log odds of −1.80 for documents with a high third topic probability and a median log odds of −2.81 for documents with a low third topic probability.

4.2.1. *Automated topic coherence*

Table 5 shows the results of the automated topic coherence analysis broken down by model and evaluator. The model fit with 100 topics has the best performance for both PCPs and students, with the model fit with 50 topics performing the worst for both groups. Example PMI scores for individual topics may be seen in Table 4.

Table 5 – Pearson correlation coefficients between model precision and PMI scores for each model and category of evaluator.

# Topics	PCP PMI	Student PMI
50	−0.17	−0.30
100	0.49	0.39
150	0.16	0.07

Please cite this article in press as: C.W. Arnold, et al., Evaluating topic model interpretability from a primary care physician perspective, Comput. Methods Programs Biomed. (2015), http://dx.doi.org/10.1016/j.cmpb.2015.10.014

5. Discussion

Although the models performed better in terms of log-likelihood with increasing numbers of topics, more topics did not translate to increased human interpretability. The 100 topic model performed significantly better than the other models with respect to the word intrusion task for both PCPs and students. These findings were reinforced by the results of the automatic coherence analysis, which determined that the 100-topic model discovered the most coherent topics. On inspection of the topics, we noticed differences in topic granularity that help to explain these results. The 50 topic model learned broad topics that at times contained common words that were not related in any obvious way. In contrast, the 150 topic model learned topics that were more focused, and grouped less-common words that together were not always distinguishable from the random intruder.

We observed that PCPs performed significantly better than students on both word and topic intrusion tasks, and required significantly less time. As students and PCPs should achieve similar results interpreting general language topics, observed differences between the groups in this study are most likely a result of the PCPs' specialized clinical knowledge and experience. Our results thus indicate that the topics learned by the models capture the specialized concepts specific to clinical documents. If plotted as a percentage of a patient's medical reports over a discrete time period (e.g., plot the average topic probability over all reports discretized by day), we believe these topics warrant further investigation for use in an automatic summarization system. In such a system, a user could review historical timelines for each topic, which would provide a PCP with a temporal orientation of medical events, and guide them to relevant documents.

A previous study by Resnik et al. presented topic model output to a clinician to evaluate the quality of topics trained on clinical text [24]. Their model was trained on Twitter data from depression patients and the resulting topics were evaluated by a clinical psychologist to determine which would be relevant in assessing a patient's level of depression. The evaluation used in their study was more targeted as the physician was looking for topics that were not only cohesive, but also related to the target medical condition. While their approach is qualitatively useful, it relies wholly on a single physician's intuition. As noted by Lau et al. [30], intuitive ratings of topics do not always correspond with the ability to perform intrusion tasks. We therefore believe that the current approach provides a more rigorous evaluation of the coherence or a topic derived from clinical text.

We found that when fit to a collection of clinical reports, LDA yielded less interpretable topics than previous experiments performed with general text collections [29]. This result may be due to the fact that topics in our work are being learned over a highly specialized collection composed of a large vocabulary of related medical terms. For example, reports detailing different surgeries can share a large number of words (e.g., *incision*, *using*, *surgeon*) and, from a bag-of-words perspective, may only be distinguishable by a small number of anatomical or procedure-specific terms. One possible way to increase the interpretability of topics could be to extend the basic LDA model to better suit the clinical reporting environment. For example, maintaining word ordering with n-grams and including structured variables that capture report contents, patient characteristics, and temporal relationships (e.g., report type, demographics, lab results, etc.) could produce topics more reflective of clinical care. While this information was not explicitly included in the current model, some of it is reflected in the text and is therefore captured in the topics generated (see supplemental material). Nevertheless, it is possible that some report metadata could improve the quality of the topics generated and prior work on general text collections could be adapted for such a model [17,39].

A limitation of this work is the relatively small sample size. Clinical reports contain information that requires specialized training to understand, as demonstrated by the significant difference in performance between students and PCPs. Because we needed to accommodate PCPs' clinical schedules, we sought to limit the amount of time required from subjects to one hour with the hope of maximizing participation while still acquiring enough data to establish significance. The topic intrusion task required interpreting reports and was therefore more time consuming, and thus fewer PCPs participated.

While the results presented here indicate that topic models can learn interpretable clinical concepts, work remains in order to translate them into a complete automatic summarization system. In a real-time system, it would be impractical to relearn topics every time a new document was added to the database. The topics, therefore, would not necessarily be optimal for new documents. For instance, if a new treatment is introduced for a given disease, it might not exist in the vocabulary of the documents used to train the model, resulting in an inability for the topics to reflect an association between the treatment and disease. A system could address this issue by periodically relearning topics using new documents, but would need to balance between consistency and plasticity. Future studies should investigate the progression of topics over time and the effect on a summarization system.

6. Conclusion

In this work we have established a baseline for the interpretability of topics learned from clinical text. While the clinical relevance of the topics was not tested, clinicians significantly outperformed lay subjects, indicating that interpretable topics capturing specialized medical information can be discovered, an important first step in utilizing them in an automatic summarization system. These results were obtained using a general topic model without any modifications for clinical reporting or additional information from clinical knowledge sources. This is an encouraging result as models tailored to clinical reporting would likely provide even greater interpretability. Our future work includes developing models that account for the various forms of data (e.g., free-text, numeric, coded) and temporal dependencies that exist in the medical record, and measuring their performance in an automatic summarization system.

Please cite this article in press as: C.W. Arnold, et al., Evaluating topic model interpretability from a primary care physician perspective, Comput. Methods Programs Biomed. (2015), http://dx.doi.org/10.1016/j.cmpb.2015.10.014

7. Competing interests statement

The authors have no competing interests to declare.

8. Funding statement

This work was supported by a grant from the National Library of Medicine of the National Institutes of Health under award number R21LM011937 (PI Arnold). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

[1] L. Pizziferri, et al., Primary care physician time utilization before and after implementation of an electronic health record: a time-motion study, J. Biomed. Inform. 38 (3) (2005) 176–188.
[2] C. Plaisant, et al., LifeLines: using visualization to enhance navigation and analysis of patient records, in: Proceedings of the AMIA Annual Symposium, Lake Buena Vista, FL, 1998.
[3] S.B. Cousins, M.G. Kahn, The visual display of temporal information, Artif. Intell. Med. 3 (6) (1991) 341–357.
[4] J.C. Feblowitz, et al., Summarization of clinical information: a conceptual model, J. Biomed. Inform. 44 (4) (2011) 688–699.
[5] Y. Shahar, et al., KNAVE-II: A distributed architecture for interactive visualization and intelligent exploration of time-oriented clinical data, in: Proceedings of Intelligent Data Analysis in Medicine and Pharmacology, Protaras, Cyprus, 2003.
[6] J.S. Hirsch, et al., HARVEST, a longitudinal patient record summarizer, J. Am. Med. Inform. Assoc. 22 (2) (2015) 263–274.
[7] T.R. Konrad, et al., It's about time: physicians' perceptions of time constraints in primary care medical practice in three national healthcare systems, Med. Care 48 (2) (2010) 95.
[8] R. Mishra, et al., Text summarization in the biomedical domain: a systematic review of recent research, J. Biomed. Inform. 52 (2014) 457–467.
[9] A. Nenkova, K. McKeown, A survey of text summarization techniques, in: Mining Text Data, Springer, 2012, pp. 43–76.
[10] R. Pivovarov, N. Elhadad, Automated methods for the summarization of electronic health records, J. Am. Med. Inform. Assoc. (2015).
[11] Y. Shahar, et al., Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions, Artif. Intell. Med. 38 (2) (2006) 115–135.
[12] W. Hsu, et al., Context-based electronic health record: toward patient specific healthcare, IEEE Trans. Inf. Technol. Biomed. 16 (2) (2012) 228–234.
[13] J.L. Boyd-Graber, D.M. Blei, X. Zhu, A Topic Model for Word Sense Disambiguation, in: Proceedings of the Joint Meeting of the Conference on Empirical Methods in Natural Language Processing and the Conference on Computational Natural Language Learning, Prague, Czech Republic, 2007.
[14] D. Blei, J. Lafferty, A correlated topic model of science, Ann. Appl. Stat. 1 (1) (2007) 17–35.
[15] D. Blei, et al., Hierarchical topic models and the nested Chinese restaurant process, in: Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference, 2003, pp. 17–24.
[16] C. Wang, D. Blei, D. Heckerman, Continuous time dynamic topic models, in: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, Helsinki, Finland, 2008, pp. 579–586.
[17] X. Wang, A. McCallum, Topics over time: a non-Markov continuous-time model of topical trends, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Philadelphia, PE, 2006, pp. 424–433.
[18] R. Cohen, et al., Redundancy-aware topic modeling for patient record notes, PLOS ONE 9 (2) (2014) pe87555.
[19] D. Newman, S. Karimi, L. Cavedon, Using topic models to interpret MEDLINE's medical subject headings, in: AI 2009: Advances in Artificial Intelligence, Springer, 2009, pp. 270–279.
[20] D. Li, et al., A bibliometric analysis on tobacco regulation investigators, BioData Min. 8 (1) (2015) 11.
[21] C. Arnold, et al., Clinical cased-based retrieval using latent topic analysis, in: Proceedings of the AMIA Annual Symposium, Washington, DC, 2010, pp. 26–31.
[22] C. Arnold, W. Speier, A topic model of clinical reports, in: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, OR, 2012.
[23] C. Howes, M. Purver, R. McCabe, Investigating topic modelling for therapy dialogue analysis, in: Proceedings IWCS Workshop on Computational Semantics in Clinical Text (CSCT), 2013.
[24] P. Resnik, et al., Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter, in: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych), 2015.
[25] Y. Halpern, et al., A comparison of dimensionality reduction techniques for unstructured clinical text, in: ICML 2012 Workshop on Clinical Data Analysis, 2012.
[26] M. Ghassemi, et al., Unfolding physiological state: mortality modelling in intensive care units, in: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, 2014.
[27] K.R. Chan, et al., An empirical analysis of topic modeling for mining cancer clinical notes, in: 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW), IEEE, 2013.
[28] T. Iwata, T. Yamada, N. Ueda, Probabilistic latent semantic visualization: topic model for visualizing documents, in: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, 2008.
[29] J. Chang, et al., Reading tea leaves: how humans interpret topic models, in: Advances in Neural Information Processing Systems 22: Proceedings of the 2009 Conference, 2009.
[30] J.H. Lau, D. Newman, T. Baldwin, Machine reading tea leaves: automatically evaluating topic coherence and topic model quality, in: Proceedings of the Association for Computational Linguistics, 2014.
[31] S.C. Deerwester, et al., Indexing by latent semantic analysis, JASIS 41 (6) (1990) 391–407.
[32] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, 1999.
[33] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. B (Methodol.) (1977) 1–38.
[34] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (5) (2003) 993–1022.
[35] H.M. Wallach, et al., Evaluation methods for topic models, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009.

Please cite this article in press as: C.W. Arnold, et al., Evaluating topic model interpretability from a primary care physician perspective, Comput. Methods Programs Biomed. (2015), http://dx.doi.org/10.1016/j.cmpb.2015.10.014

[36] M. Buhrmester, T. Kwang, S.D. Gosling, Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? Perspect. Psychol. Sci. 6 (1) (2011) 3–5.
[37] A.K. McCallum, MALLET: A Machine Learning for Language Toolkit, 2002.
[38] D. Newman, et al., Automatic evaluation of topic coherence, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2010.
[39] D. Mimno, A. McCallum, Topic models conditioned on arbitrary features with dirichlet-multinomial regression, in: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, Helsinki, Finland, 2008, pp. 411–418.

Please cite this article in press as: C.W. Arnold, et al., Evaluating topic model interpretability from a primary care physician perspective, Comput. Methods Programs Biomed. (2015), http://dx.doi.org/10.1016/j.cmpb.2015.10.014