A Topic Model of Clinical Reports

Corey Arnold, PhD and William Speier, MS
Medical Imaging Informatics Group, UCLA
924 Westwood Blvd Ste 420
Los Angeles, CA 90024
cwarnold@ucla.edu, speier@mii.ucla.edu

ABSTRACT
Clinical narrative in the medical record provides perhaps the most detailed account of a patient's history. However, this information is documented in free-text, which makes it challenging to analyze. Efforts to index unstructured clinical narrative often focus on identifying predefined concepts from clinical terminologies. Less studied is the problem of analyzing the text as a whole to create temporal indices that capture relationships between learned clinical events. Topic models provide a method for analyzing large corpora of text to discover semantically related clusters of words. This work presents a topic model tailored to the clinical reporting environment that allows for individual patient timelines. Results show the model is able to identify patterns of clinical events in a cohort of brain cancer patients.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Information filtering.

General Terms
Algorithms.

Keywords
Topic model, clinical reports, brain cancer.

1. INTRODUCTION
Due to the complexity of language, the variability in author reporting styles, and differences in clinical practice, clinical narrative can be challenging to analyze from a computer's perspective. Topic models, such as latent Dirichlet allocation (LDA), provide a method for indexing large unstructured corpora with inferred semantics [1]. Extensions to the LDA model have been proposed that include modeling time [2], finding correlations between topics [3], learning image-word annotations [4], performing automatic translation [5], and learning topic hierarchies [6]. Additionally, previous work has demonstrated the application of LDA in the clinical domain for case-based reasoning [7]. However, to date there is no model designed specifically for clinical reporting, where each patient has a collection of documents that details the progression of disease. We propose a topic model that captures temporal topic patterns in an individual patient's medical record, while being sensitive to the entire patient population.

2. MODEL DESIGN
Similar to [2], we present a model that links topics to time within a corpus by observing a timestamp for each document and using a beta distribution to model a topic's expression in the collection over time. However, we modify the model for application in the clinical world, where each patient has his or her own timeline that displays only a subset of all possible topics. We introduce a patient plate *P* to account for variations in topic expression over time between patients. Figure 1 illustrates this extension, denoted by the outermost box, which contains the time modeling beta distributions, π. As denoted by the topic plate *T*, which falls outside of the patient plate, topics are learned from the entire collection of patient documents. This design fits with our expectation that for a cohort of patients there is a superset of topics, with each patient expressing a subset of topics specific to their individual disease progression. Furthermore, we expect temporal relationships between subsets of topics to generalize across patients (e.g., a patient must first undergo a surgical resection of a tumor before a radiologist notes the resection cavity in a subsequent MRI study). We define the following generative process for the model:

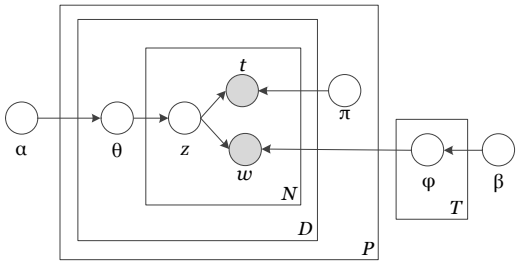

Figure 1: A topic model of clinical reports.

1. For each topic draw $\varphi_t \sim \text{Dirichlet}(\beta)$
2. For each patient *p*
   a. For each document *d* from patient *p* draw $\theta_{pd} \sim \text{Dirichlet}(\alpha)$
   b. For each word *i* in document *pd*
      i. Draw a topic $z_{pdi} \sim Multinomial(\theta_{pd})$
      ii. Draw a word $w_{pdi} \sim \text{Multinomial}(\varphi_{z_{pdi}})$
      iii. Draw a timestamp $t_{pdi} \sim \text{Beta}(\pi_{z_{pdi}})$

We fit the model parameters using a Gibbs sampling procedure that calculates the conditional probability of a topic as

$$P(z_{pdi} = j \mid \mathbf{w}, \mathbf{t}, \mathbf{z}_{\neg pdi}, \alpha, \beta, \pi_p) \propto \frac{n_{\neg i,j}^{(w_i)} + \beta - 1}{n_{\neg i,j}^{(\bullet)} + W\beta - 1} \frac{n_{\neg i,j}^{(pd_i)} + \alpha - 1}{n_{\neg i,\bullet}^{(pd_i)} + T\alpha - 1} \frac{(1 - t_{pdi})^{\pi_{z_{pdi}1} - 1} t_{pdi}^{\pi_{z_{pdi}2} - 1}}{B(\pi_{z_{pdi}1}, \pi_{z_{pdi}2})}$$

where $n_{\neg i}^{(\bullet)}$ is a count that does not include the assignment for the current word, z_{pdi}, and $B(\,)$ is the beta function. The parameters for a patient's topic beta distributions (π_1 and π_2) are updated after each Gibbs sample using the method of moments.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

Table 1: Example topics learned by the model.

Topic	Label
left brain mass tumor temporal lobe mri contrast cm frontal	"imaging diagnosis"
tumor resection craniotomy head area left flap intraoperative incision scalp	"surgical resection"
radiation treatment therapy cgy total dose physician site oncology outpatient	"radiation treatment"
brain enhancement axial mri contrast cavity resection signal prior scan	"post treatment imaging surveillance"

3. RESULTS

We investigated the use of the proposed topic model in a population of patients with glioblastoma multiforme (GBM), an aggressive brain cancer. For each patient we used any report that conveyed clinical information in natural language (e.g., discharge summaries, radiology reports, pathology reports, etc.). Patients were required to have a minimum of five reports to be included in the model. Reports were preprocessed to remove stop words, rare and common words, and a set of medical stop words, such as "Dr.", "report", "dictated", and "ID". In a general context these medical stop words may prove useful, but limited to the domain of clinical reporting they offer little semantic meaning. The resulting corpus contained 303 patients, 13,028 reports, 2,412,385 words, and 1,374 unique words. Each patient's document collection was normalized to the timespan (0,1) and 100 topics were fit in 1000 iterations. As in [2], we used symmetric Dirichlet priors of $\alpha = 50/T$ and $\beta = 0.1$.

The resulting topics and temporal patterns were reviewed by a neuroradiologist and were found to correlate with valid sequences of clinical events. For example, we observed that the topic describing radiation treatment is generally preceded by the topic describing the surgical resection of a tumor. Table 1 presents several topics learned by the model and Figure 2 shows the topic timeline of a patient from the collection. As expected, because topics are learned across all patients, we found that generally patients exhibit only a subset of all possible topics. For example, some patients have tumors considered inoperable and therefore do not express a surgical resection topic.

4. DISCUSSION

Patients with large numbers of documents can bias the distribution of words in a topic. With a large enough collection of patients these biases may even out, but it is likely that the differences in numbers of documents per patient are systematic (e.g., patients with newly discovered late stage tumors may die quickly and therefore have few documents). This may be corrected by estimating topics through sampling word topic counts from patients as a proportion of their number of documents.

We observed that while general temporal trends were found, there were cases of topic expression that conflicted with actual patient care. For example, we identified patients who began physical therapy before receiving any treatment for a tumor. This discrepancy is due to the fact that although relative temporal topic relationships hold across patients, they are expressed at different points within each patient's timeline. Additionally, due to the nature of the clinical corpus, where similar language is used across reports, this problem is compounded as there are a small number of unique words relative to the total number of words.

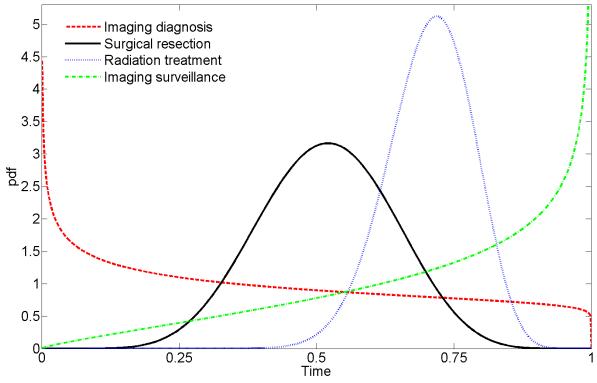

Figure 2: Beta probability density functions (pdf) for four topics in a patient with 48 documents spanning three years.

5. CONCLUSION

The proposed model was able to learn temporal trends between topics in patient records. These topics may be useful for identifying patients based on patterns of temporal-topic expression for predictive purposes or for cohort finding. Topics may also be used for case-based reasoning by comparing topic-time distributions across patients. A metric, such as Kullback-Liebler divergence, may be used for this task. Future work includes pursuing these applications and defining model configurations that allow for more flexible individual topic expression.

6. ACKNOWLEDGEMENTS

This work was supported by R01-LM009961 and T15-LM07536.

7. REFERENCES

[1] Blei, D., Ng, A. and Jordan, M. Latent Dirichlet allocation. *J. of Mach. Learn. Res.* 3 (Jan. 2003), 993-1022.

[2] Wang, X. and McCallum, A. Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the SIGKDD Conference on Knowledge Discovery and Data Mining* (Philadelphia, PA, Aug. 20-23, 2006). SIGKDD '06. ACM New York, NY, 424-433.

[3] Blei, D. and Lafferty, J. A correlated topic model of science. *Ann. Appl. Stat.* 1,1 (Jun. 2007), 17-35.

[4] Blei, D. and Jordan, M. Modeling annotated data. In *Proceedings of the SIGIR Conference on Research and Development in Information Retrieval* (Toronto, Canada, Jul. 28-Aug. 1, 2003). SIGIR '03. ACM New York, NY, 127-134.

[5] Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A. and McCallum, A. Polylingual topic models. In *Proceedings of the 14th Conference on Empirical Methods in Natural Language Processing* (Singapore, 2009).

[6] Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. Hierarchical dirichlet processes. *J. A. Stat. Assoc.* 101,476 (Dec. 2006), 1566-1581.

[7] Arnold C., El-Saden, S., Bui A. and Taira, R. Clinical case-based retrieval using latent topic analysis. In *Proceedings of the American Medical Informatics Association Annual Symposium* (Nov. 13-17, 2010), AMIA '10, 26-30.