Dialect Topic Modeling for Improved Consumer Medical Search

Steven P. Crain, BS,[a] Shuang-Hong Yang, MS,[a] Hongyuan Zha, PhD,[a] Yu Jiao, PhD[b]

[a]Georgia Institute of Technology, Atlanta, GA
[b]Oak Ridge National Laboratory, Oak Ridge, TN

Abstract

Access to health information by consumers is hampered by a fundamental language gap. Current attempts to close the gap leverage consumer oriented health information, which does not, however, have good coverage of slang medical terminology. In this paper, we present a Bayesian model to automatically align documents with different dialects (slang, common and technical) while extracting their semantic topics. The proposed diaTM model enables effective information retrieval, even when the query contains slang words, by explicitly modeling the mixtures of dialects in documents and the joint influence of dialects and topics on word selection. Simulations using consumer questions to retrieve medical information from a corpus of medical documents show that diaTM achieves a 25% improvement in information retrieval relevance by nDCG@5 over an LDA baseline.

1 Introduction

The Internet makes a wealth of health information available to consumers, ranging from consumer-oriented resources like WebMD[2] to technical journal articles available through PubMed Central[1]. However, many users do not know enough about their problem and the relevant technical language to form an appropriate technical query[3]. Instead, many pose the question on question answering sites, where they can use familiar language. For example, one user on Yahoo! Answers asked why his eyelids sometimes beat uncontrollably. The user was describing a mild case of blepharospasm, but, without knowing how to describe the behavior in technical language, the user could not find any relevant information. In many cases the questions are couched in slang language, like "gooey" or "preggers" (pregnant). A system that could locate health-related information from a common language or slang query would greatly benefit consumers.

In this work, we show how to accommodate dialects (variations within a language) when retrieving health information using Dialect Topic Models (diaTM). DiaTM automatically learns a set of topics and how they are expressed in several dialects. By comparing documents by topic across dialects, diaTM can find relevant technical documents for a user's non-technical query. It can also help filter out incomprehensible documents based on its assessment of the dialect gap between the user and the document. Although we cannot replace the human element in question answering services, we can help consumers with typical health literacy find the health resources they need.

2 Related Work

Many researchers grappled with the language gap between consumers and medical documents. Zeng et al.[4] showed that this gap substantially degrades search ability and satisfaction. HIQuA expanded the user's original query with technical words semantically close to the words in the query[5], but the researchers found no improvement in user ability to find necessary information. MedicoPort[6] was based on the hypothesis that co-occurrence in WebMD® would enable more useful query expansion. Users were able to find many more relevant documents, but only when the query terms were present in WebMD®. MedSearch[7] attempts to tackle the problem by accepting longer queries and distilling them to shorter, more technical queries. It also uses clustering to increase the variety of search results. Our unique contribution is that we incorporate dialect into the model so that we can perform a topical comparison of a non-technical query and a technical document, without going through query expansion. We use term frequency statistics from various collections of documents to provide dialect information to the model, which Keselman et al. found correlated strongly with consumer comprehension[8].

DiaTM is a variation on Latent Dirichlet Allocation (LDA)[9]. LDA assumes that each document is woven from a set of topics, and that the probability of words appearing in the document is related to these topics. So, for example, a document about cancer is more likely to contain "chemotherapy" than a document about pregnancy. A collection of word probabilities, like that used by LDA, is called a language model (LM). LDA uses one LM for each topic, so that the LM of a document is a mixture of the topic LMs.

Polylingual Topic Models (pTM)[10] is an extension of LDA that allows cross-language correlation of topics. By processing special pairs of documents with similar topics in different languages, pTM learns a LM in each language for each topic. So, where LDA might learn separate topics for pregnancy in each language, pTM will learn a single pregnancy topic with language-specific language models. That allows it to measure the topical similarity between two documents, even though they are in different languages. However, pTM requires parallel documents and clear labeling of languages that makes it impractical for medical dialects. Our model directly solves these two issues.

3 Dialectical Latent Dirichlet Allocation

Consumer medical information retrieval presents several unique challenges. First, documents are seldom written in a single dialect. Consider the slang sentence, "Since that time my eyes are always oozing green yellow gunk." This is a mixture of slang words ("oozing," "gunk" and "wont") and many words that could appear equally well in a technical journal paper. The more serious limitation is that training polylingual topic models requires parallel documents containing the same topics in different languages. Resources like Wikipedia provide suitable parallel documents in multiple languages, but there is no natural source of parallel slang and technical documents. Together, these characteristics make reasoning about topics in languages of several dialects very challenging.

We propose Dialect Topic Models (diaTM) to address these difficulties. As in pTM, topic selection is independent of dialect but the dialect influences the words used to express a topic. The key differences are that the dialect is unknown (must be inferred) and that it may be different for each word in the document. To help the model determine the dialect of each word, we provide features that are chosen to have good correlation with one or more dialects. For example, a technical term is expected to occur more frequently than a slang word in technical contexts, so that word frequencies in largely technical or slang collections of documents is a useful feature[8].

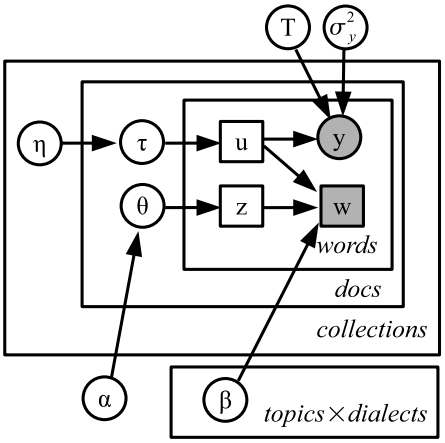


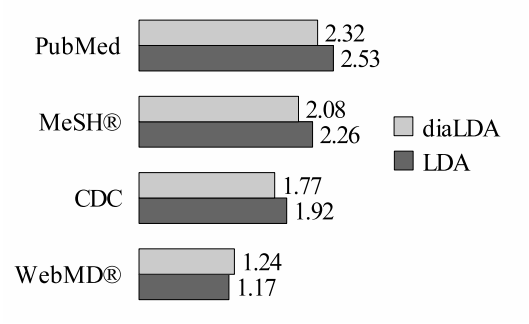
Figure 1: Graphical model of diaTM.

The structure of diaTM is shown in Figure 1. As is typical of LM approaches, the model assumes that documents are generated randomly, even though in reality careful thought is behind them. This allows us to measure how well the model explains a set of documents (Section 4 Learning Topics). For each document, a multinomial mixture of topics θ is chosen from a Dirichlet distribution that uses the same parameter α for every topic, so that on the whole every topic is equally represented. Similarly, a multinomial mixture of dialects τ is chosen from a Dirichlet distribution, but the parameters η_{cu} vary both by collection c and by dialect u, thus allowing the dialects to have different biases in each collection. For each word, a topic z and a dialect u are selected from corresponding mixtures. Then, the word itself w is selected from a multinomial distribution β_{zu} that is specific to the selected topic and dialect. Finally, the vector of dialect features y is selected by applying a linear transformation T to the expected word dialect and adding Gaussian noise with variance σ_y^2.

This model can be learned using documents from a number of different sources that have different dialect mixtures. Following Blei et al.[9], we use variational approximation to learn the model and to analyze new documents. Rather than learning T, it is expedient to learn T^{-1} by ordinary least squares regression[11]. This allows us to infer the dialect from the features, $u = T^{-1}y - \epsilon$.

Once the model has been trained, diaTM can analyze new documents. Given the document, its source and the word features, diaTM can infer the expected mixture of topics θ and dialects τ of the document and the most likely topic z and dialect u of each word. θ can be used to compare a query q and a document d

(a) Perplexity.



(b) KL-Divergence.

Figure 2: DiaTM outperforms the LDA baseline using two important metrics for which lower scores are better.

using cosine distance[12], $\theta_q \cdot \theta_d / ||\theta_q|| ||\theta_d||$. This distance depends on the topics present in the query and document, but does not depend on the dialects. The model can also measure the distance between two dialect mixtures τ, in order to gauge whether the user will understand a given document. Finally, the model predicts the dialect u of each word in a query, which would be valuable in a large-scale medical information retrieval system to apply heterogeneous search strategies tuned for specific dialects.

4 Experimental Results

Data Sets. We collected documents with different mixtures of dialects: common with some slang (Yahoo! Answers health category[13]); technical (PubMed Central Open Access Subset[14] and Medical Subject Headings[15] (MeSH®)); and common with technical (WebMD®[2] and Centers for Disease Control and Prevention[16] Websites). For each collection, we randomly selected 1000 English-language documents for training and 2500 each for validation (by perplexity) and testing. We stemmed the tokens with the SnowBall Stemmer[17] and then selected the 2000 words occurring in the most documents, excluding a hand-selected list of about 300 uninformative words This resulted in combined vocabulary containing 4778 words.

To provide a signal for identifying word dialect, we used the term frequency and document frequency of the word in each collection as a feature. Because we separated the Yahoo! questions and answers, this resulted in 12 features. The thirteenth feature was context dependent: the fraction of the occurrences of the word within the document that appeared in the question portion of the document. This was of course zero in every collection except Yahoo!, where it ranges from 0 (only in answer) to 1 (only in question).

Loose weight fast!
I weigh 163 and i'm 5"6. I need to drop pounds as quickly as possible, I want to loose about 25 pounds or more, I do situps every evening and I eat mostly veggies and lean meat. Please help!!

(a) Question (modified to protect privacy).

Skipping breakfast is even **harder** on your brain. Most of the cells in your body can store energy up for lean times, but your brain cells need a constant supply of carbohydrates to function.

(b) LDA[18].

If eating cabbage soup for a solid week appeals to you, the Cabbage Soup **Diet** is sure to lead to quick weight loss. However, since the *food* **choices** are so **limited** and the calories so **low**, boredom—and inadequate **nutrition**—are inevitable.

Dialects: omitted from vocab; consumer; **common**; *technical*.

(c) DiaTM[19].

Figure 3: Given a consumer's question (a), LDA identifies a best matching document (b) that is not as directly on topic as the document found by diaTM (c).

Evaluation of Learning Topics. DiaTM learns to identify the topics that are present in documents, using characteristic probability distributions over words. Perplexity is commonly used to compare language models[20]. If the model does a good job of explaining the words in a test set, it will have low perplexity. DiaTM had 54.7% better perplexity than LDA (Figure 2(a)).

$$pplex = \exp\left(\sum_j P(j|testset)\log(P(j|model))\right) \quad (1)$$

Topical Comparison. We use Kullback-Leibler (KL) divergence to measure the extent to which technical and non-technical documents use the same topics. For a given collection c and diaTM model, it is straight-forward to compute the conditional probability distribution over topics z, from which we can compute the KL divergence. A model has a better mix of

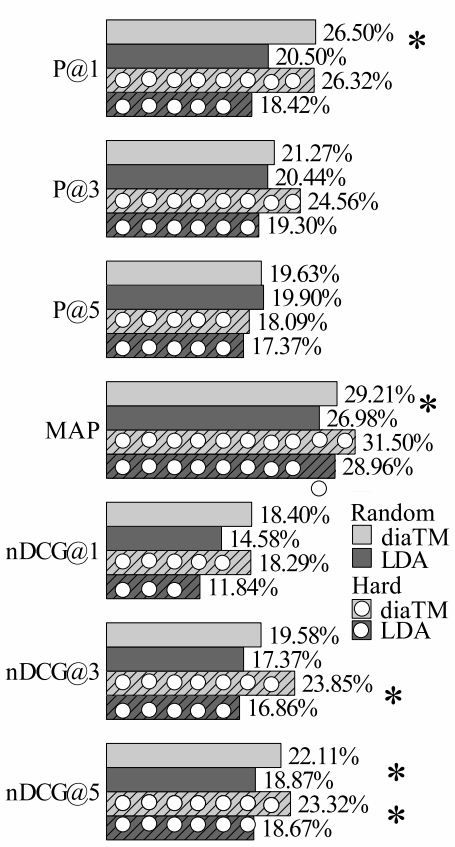


Figure 4: DiaTM outperforms the LDA baseline on the IR task. Performance shown for general and hard (heavily slang) queries. Higher scores mean better performance. * Significant at p=0.04.

collections (and presumably of dialects) if the KL divergences are small. On this measure, diaTM again outperforms the baseline by about 8%, as shown in Figure 2(b).

$$KL_c = \sum_z P(z|\text{Yahoo!}) \log \left(\frac{P(z|\text{Yahoo!})}{P(z|c)} \right) \quad (2)$$

Effectiveness for Information Retrieval. We selected 38 questions from the test set that contained substantial slang content and another 162 questions randomly. For each question, we selected documents that were highly ranked by one of many models, not including either model evaluated here. We obtained the judgments using Mechanical Turk: editors were asked to assign one of five grades based on the topical similarity. The high relevance grades were audited, and the highest approved rating for each pair was used. In this way we obtained 5854 judgments on 4982 query-document pairs, which we are making available to other researchers[21].

Figure 3 shows the highest ranked document identified by each algorithm, using topical cosine similarity. Notice that diaTM picked a document on the same topic (weight loss diet) whereas LDA picked a document on a related topic (nutrition). We evaluated the performance using standard metrics implemented in the LETOR[22] evaluation tools: normalized Discounted Cumulative Gain (nDCG), Precision (P) and Mean Average Precision (MAP). The results are shown in Figure 4. DiaTM outperformed LDA on most of the metrics, including a 24.8% improvement of nDCG@5 for the harder slang queries.

5 Discussion

Many researchers have grappled with the gap between the language of consumer problems and the technical language of the documents they are looking for, yet this remains a largely unsolved problem. In this paper we have described diaTM, an extension of pTM to handle the unique characteristics of this language gap.

In comparison to LDA, a state of the art model for topical inference in large collections of documents, diaTM has less than half the perplexity. That is a tremendous improvement in perplexity, and indicates that the dialect information that we incorporate is very valuable for understanding the documents in an abstract sense. Moreover, the improvement in information retrieval performance is also dramatic. DiaTM holds great promise for improving consumer medical search.

On the other hand, there are still a number of weaknesses in diaTM. The improvement in KL divergence is a modest 8%, so that there is not the substantial sharing of topics that we expected. Additionally, the "dialects" that the model learned were tied quite strongly to the features and do not reflect the slang and technical dialects very well. The strong mismatch between topics in the technical collections and in the consumer questions seems to be largely to blame. For example, gene sequencing and protein structure were common technical topics but completely absent from the consumer questions. Also, the vocabulary selection technique we used did not provide a good enough coverage of slang words, which are individually rare even in consumer questions. DiaTM may fare even better when trained with a richer slang vocabulary and a larger collection of slang and common documents.

In summary, diaTM results in substantial improvements in retrieval performance using real user queries, improvements that are sorely needed.

Acknowledgments

This research was supported by an Oak Ridge Computational Science and Engineering Fellowship. It was also partially supported by grants from Microsoft and Hewlett-Packard. We thank the reviewers for much helpful feedback.

References

[1] National Institutes of Health. PubMed central: A free archive of life sciences journals; 2009. Available from: http://www.ncbi.nlm.nih.gov/pmc.

[2] WebMD. WebMD®: Better information. Better health; 2009. Available from: http://www.webmd.com.

[3] Chan CV, Matthews LA, Kaufman DR. A taxonomy characterizing complexity of consumer ehealth literacy. In: AMIA Annual Symposium Proceedings. vol. 2009; 2009. p. 86–90.

[4] Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA. Characteristics of consumer terminology for health information retrieval. Methods Inf Med. 2002;41:289–298.

[5] Zeng QT, Crowell J, Plovnick RM, Kim E, Ngo L, Dibble E. Assisting consumer health information retrieval with query recommendations. Journal of the American Medical Informatics Association. 2006;13(1):80 – 90.

[6] Can AB, Baykal N. MedicoPort: A medical search engine for all. Computer Methods and Programs in Biomedicine. 2007;86(1):73 – 86.

[7] Luo G, Tang C, Yang H, Wei X. MedSearch: a specialized search engine for medical information retrieval. In: CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management; 2008. p. 143–152.

[8] Keselman A, Tse T, Crowell J, Browne A, Ngo L, Zeng Q. Assessing Consumer Health Vocabulary Familiarity: An Exploratory Study. Journal of Medical Internet Research. 2007;9(1).

[9] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.

[10] Mimno D, Wallach HM, Naradowsky J, Smith DA, McCallum A. Polylingual topic models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing; 2009. p. 880–889.

[11] Farebrother RW. Linear least squares computations. Marcel Dekker, New York; 1987.

[12] Frakes WB, Baeza-Yates R, editors. Information retrieval, data structure and algorithms. Prentice Hall; 1992.

[13] Yahoo!. Yahoo! Answers health category; 2010. RSS feed. Available from: http://answers.yahoo.com/rss/catq?sid=396545018.

[14] National Institutes of Health. PubMed Central open access subset; 2009. Available from: ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/articles.tar.gz.

[15] U S National Library of Medicine. Medical subject headings descriptors; 2010. Available from: http://www.nlm.nih.gov/mesh/filelist.html.

[16] Centers for Disease Control and Prevention. Centers for Diseas Control and Prevention: Your online source for credible health information; 2009. Available from: http://www.cdc.gov.

[17] Porter M. Snowball; 2009. Software. Available from: http://snowball.tartarus.org.

[18] Sullivan D. Eat right for energy. In: Fitness & Nutrition. Blue Cross Blue Shield of Massachusetts; 2009. Available from: http://www.ahealthyme.com/topic/dietenergy.

[19] Zelman KM. The cabbage soup diet. In: Healthy eating & diet. WebMD; 2008. Available from: http://www.webmd.com/diet/features/the-cabbage-soup-diet.

[20] Bahl L, Baker J, Jelinek E, Mercer R. Perplexity—a measure of the difficulty of speech recognition tasks. In: Program, 94th Meeting of the Acoustical Society of America. vol. 62; 1977. p. S63.

[21] Crain SP, Zha H. Consumer Medical Information Retrieval Relevance Judgments; 2010. Available from: https://research.cc.gatech.edu/dmirlab/node/2.

[22] Qin T, Liu TY, Xu J, Li H. LETOR: A benchmark collection for research on learning to rank for information retrieval. Information Retrieval Journal. 2010;Available from: http://research.microsoft.com/en-us/um/beijing/projects/letor/.