Generative Content Models for Structural Analysis of Medical Abstracts

Jimmy Lin\textsuperscript{1,2}, Damianos Karakos\textsuperscript{3}, Dina Demner-Fushman\textsuperscript{2}, and Sanjeev Khudanpur\textsuperscript{3}

\textsuperscript{1}College of Information Studies  \hspace{1cm}  \textsuperscript{3}Center for Language and Speech Processing
\textsuperscript{2}Institute for Advanced Computer Studies  \hspace{1cm}  Johns Hopkins University
University of Maryland  \hspace{1cm}  College Park, MD 20742, USA  \hspace{1cm}  Baltimore, MD 21218, USA
jimmylin@umd.edu, demner@cs.umd.edu (damianos, khudanpur)@jhu.edu

Abstract
The ability to accurately model the content structure of text is important for many natural language processing applications. This paper describes experiments with generative models for analyzing the discourse structure of medical abstracts, which generally follow the pattern of “introduction”, “methods”, “results”, and “conclusions”. We demonstrate that Hidden Markov Models are capable of accurately capturing the structure of such texts, and can achieve classification accuracy comparable to that of discriminative techniques. In addition, generative approaches provide advantages that may make them preferable to discriminative techniques such as Support Vector Machines under certain conditions. Our work makes two contributions: at the application level, we report good performance on an interesting task in an important domain; more generally, our results contribute to an ongoing discussion regarding the tradeoffs between generative and discriminative techniques.

1 Introduction
Certain types of text follow a predictable structure, the knowledge of which would be useful in many natural language processing applications. As an example, scientific abstracts across many different fields generally follow the pattern of “introduction”, “methods”, “results”, and “conclusions” (Salanger-Meyer, 1990; Swales, 1990; Orăsan, 2001). The ability to explicitly identify these sections in unstructured text could play an important role in applications such as document summarization (Teufel and Moens, 2000), information retrieval (Tbahriti et al., 2005), information extraction (Mizuta et al., 2005), and question answering. Although there is a trend towards analysis of full article texts, we believe that abstracts still provide a tremendous amount of information, and much value can still be extracted from them. For example, Gay et al. (2005) experimented with abstracts and full article texts in the task of automatically generating index term recommendations and discovered that using full article texts yields at most a 7.4% improvement in F-score. Demner-Fushman et al. (2005) found a correlation between the quality and strength of clinical conclusions in the full article texts and abstracts.

This paper presents experiments with generative content models for analyzing the discourse structure of medical abstracts, which has been confirmed to follow the four-section pattern discussed above (Salanger-Meyer, 1990). For a variety of reasons, medicine is an interesting domain of research. The need for information systems to support physicians at the point of care has been well studied (Covell et al., 1985; Gorman et al., 1994; Ely et al., 2005). Retrieval techniques can have a large impact on how physicians access and leverage clinical evidence. Information that satisfies physicians' needs can be found in the MEDLINE database maintained by the U.S. National Library of Medicine.
(NLM), which also serves as a readily available corpus of abstracts for our experiments. Furthermore, the availability of rich ontological resources, in the form of the Unified Medical Language System (UMLS) (Lindberg et al., 1993), and the availability of software that leverages this knowledge—MetaMap (Aronson, 2001) for concept identification and SemRep (Rindflesch and Fiszman, 2003) for relation extraction—provide a foundation for studying the role of semantics in various tasks.

McKnight and Srinivasan (2003) have previously examined the task of categorizing sentences in medical abstracts using supervised discriminative machine learning techniques. Building on the work of Ruch et al. (2003) in the same domain, we present a generative approach that attempts to directly model the discourse structure of MEDLINE abstracts using Hidden Markov Models (HMMs); cf. (Barzilay and Lee, 2004). Although our results were not obtained from the same exact collection as those used by authors of these two previous studies, comparable experiments suggest that our techniques are competitive in terms of performance, and may offer additional advantages as well.

Discriminative approaches (especially SVMs) have been shown to be very effective for many supervised classification tasks; see, for example, (Joachims, 1998; Ng and Jordan, 2001). However, their high computational complexity (quadratic in the number of training samples) renders them prohibitive for massive data processing. Under certain conditions, generative approaches with linear complexity are preferable, even if their performance is lower than that which can be achieved through discriminative training. Since HMMs are very well-suited to modeling sequences, our discourse modeling task lends itself naturally to this particular generative approach. In fact, we demonstrate that HMMs are competitive with SVMs, with the added advantage of lower computational complexity. In addition, generative models can be directly applied to tackle certain classes of problems, such as sentence ordering, in ways that discriminative approaches cannot readily. In the context of machine learning, we see our work as contributing to the ongoing debate between generative and discriminative approaches—we provide a case study in an interesting domain that begins to explore some of these tradeoffs.

2 Methods
2.1 Corpus and Data Preparation

Our experiments involved MEDLINE, the bibliographical database of biomedical articles maintained by the U.S. National Library of Medicine (NLM). We used the subset of MEDLINE that was extracted for the TREC 2004 Genomics Track, consisting of citations from 1994 to 2003. In total, 4,591,008 records (abstract text and associated metadata) were extracted using the Date Completed (DCOM) field for all references in the range of 19940101 to 20031231.

Viewing structural modeling of medical abstracts as a sentence classification task, we leveraged the existence of so-called structured abstracts (see Figure 1 for an example) in order to obtain the appropriate section label for each sentence. The use of section headings is a device recommended by the Ad Hoc Working Group for Critical Appraisal of the Medical Literature (1987) to help humans assess the reliability and content of a publication and to facilitate the indexing and retrieval processes. Although structured abstracts loosely adhere to the introduction, methods, results, and conclusions format, the exact choice of section headings varies from abstract to abstract and from journal to journal. In our test collection, we observed a total of 2688 unique section headings in structured abstracts—these were manually mapped to the four broad classes of “introduction”, “methods”, “results”, and “conclusions”. All sentences falling under a section heading were assigned the label of its appropriately-mapped heading (naturally, the actual section headings were removed in our test collection). As a concrete example, in the abstract shown in Figure 1, the “OBJECTIVE” section would be mapped to “introduction”, the “RESEARCH DESIGN AND METHODS” section to “methods”. The “RESULTS” and “CONCLUSIONS” sections map directly to our own labels. In total, 308,055 structured abstracts were extracted and prepared in this manner, serving as the complete dataset. In addition, we created a reduced collection of 27,075 abstracts consisting of only Randomized Controlled Trials (RCTs), which represent definitive sources of evidence highly-valued in the clinical decision-making process.

Separately, we manually annotated 49 unstruc-
Integrating medical management with diabetes self-management training: a randomized control trial of the Diabetes Outpatient Intensive Treatment program.

OBJECTIVE– This study evaluated the Diabetes Outpatient Intensive Treatment (DOIT) program, a multiday group education and skills training experience combined with daily medical management, followed by case management over 6 months. Using a randomized control design, the study explored how DOIT affected glycemic control and self-care behaviors over a short term. The impact of two additional factors on clinical outcomes were also examined (frequency of case management contacts and whether or not insulin was started during the program).

RESEARCH DESIGN AND METHODS– Patients with type 1 and type 2 diabetes in poor glycemic control (A1c $\geq$8.5%) were randomly assigned to DOIT or a second condition, entitled EDUPOST, which was standard diabetes care with the addition of quarterly educational mailings. A total of 167 patients (78 EDUPOST, 89 DOIT) completed all baseline measures, including A1c and a questionnaire assessing diabetes-related self-care behaviors. At 6 months, 117 patients (52 EDUPOST, 65 DOIT) returned to complete a follow-up A1c and the identical self-care questionnaire.

RESULTS– At follow-up, DOIT evidenced a significantly greater drop in A1c than EDUPOST. DOIT patients also reported significantly more frequent blood glucose monitoring and greater attention to carbohydrate and fat contents (ACFC) of food compared with EDUPOST patients. An increase in ACFC over the 6-month period was associated with improved glycemic control among DOIT patients. Also, the frequency of nurse case manager follow-up contacts was positively linked to better A1c outcomes. The addition of insulin did not appear to be a significant contributor to glycemic change.

CONCLUSIONS– DOIT appears to be effective in promoting better diabetes care and positively influencing glycemia and diabetes-related self-care behaviors. However, it demands significant time, commitment, and careful coordination with many health care professionals. The role of the nurse case manager in providing ongoing follow-up contact seems important.

Figure 1: Sample structured abstract from MEDLINE.
In an attempt to further boost performance, we employed Linear Discriminant Analysis (LDA) to find a linear projection of the four-dimensional vectors that maximizes the separation of the Gaussians (corresponding to the HMM states). Venables and Ripley (1994) describe an efficient algorithm (of linear complexity in the number of training sentences) for computing the LDA transform matrix, which entails computing the within- and between-covariance matrices of the classes, and using Singular Value Decomposition (SVD) to compute the eigenvectors of the new space. Each sentence/vector is then multiplied by this matrix, and new HMM models are re-computed from the projected data.

An important aspect of our work is modeling content structure using generative techniques. To assess the impact of taking discourse transitions into account, we compare our fully trained model to one that does not take advantage of the Markov assumption—i.e., it assumes that the labels are independently and identically distributed.

To facilitate comparison with previous work, we also experimented with binary classifiers specifically tuned to each section. This was done by creating a two-state HMM: one state corresponds to the label we want to detect, and the other state corresponds to all the other labels. We built four such classifiers, one for each section, and trained them in the same manner as above.

### 3 Results

We report results on three distinct sets of experiments: (1) ten-fold cross-validation (90/10 split) on all structured abstracts from the TREC 2004 MEDLINE corpus, (2) ten-fold cross-validation (90/10 split) on the RCT subset of structured abstracts from the TREC 2004 MEDLINE corpus, (3) training on the RCT subset of the TREC 2004 MEDLINE corpus and testing on the 49 hand-annotated held-out testset.

The results of our first set of experiments are shown in Tables 1(a) and 1(b). Table 1(a) reports the multi-way classification error rate; once again, applying the Markov assumption to model discourse transitions improves performance, and using LDA further reduces error rate. Table 2(b) reports accuracy, precision, recall, and F-measure for four separate binary classifiers specifically trained for each of the sections (one per row in the table). We only display results with our best model, namely HMM with LDA.

The results of our second set of experiments (with RCTs only) are shown in Tables 2(a) and 2(b). Table 2(a) reports the multi-way classification error rate; once again, applying the Markov assumption to model discourse transitions improves performance, and using LDA further reduces error rate. Table 2(b) reports accuracy, precision, recall, and F-measure for four separate binary classifiers (HMM with LDA) specifically trained for each of the sections (one per row in the table). The table also presents the closest comparable experimental results reported by McKnight and Srinivasan (2003).\(^1\)

---

\(^1\)After contacting the authors, we were unable to obtain the same exact dataset that they used for their experiments.
Table 2: Ten-fold cross-validation results on the structured RCT subset of the TREC 2004 MEDLINE corpus: multi-way classification (a) and binary classification (b). Table (b) also reproduces the results from McKnight and Srinivasan (2003) for a comparable task on a different RCT-subset of structured abstracts.

(a)

<table>
<thead>
<tr>
<th>Model</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-HMM</td>
<td>.238</td>
</tr>
<tr>
<td>HMM</td>
<td>.212</td>
</tr>
<tr>
<td>HMM + LDA</td>
<td>.209</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Section</th>
<th>Present study</th>
<th>McKnight and Srinivasan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>Prec</td>
</tr>
<tr>
<td>Introduction</td>
<td>.931</td>
<td>.898</td>
</tr>
<tr>
<td>Methods</td>
<td>.904</td>
<td>.812</td>
</tr>
<tr>
<td>Results</td>
<td>.902</td>
<td>.902</td>
</tr>
<tr>
<td>Conclusions</td>
<td>.929</td>
<td>.772</td>
</tr>
</tbody>
</table>

Table 3: Training on the structured RCT subset of the TREC 2004 MEDLINE corpus, testing on corpus of hand-annotated abstracts: multi-way classification (a) and binary classification (b). Unstructured abstracts with all four sections (complete), and with missing sections (partial) are shown. Table (b) again reproduces the results from McKnight and Srinivasan (2003) for a comparable task on a different subset of 206 unstructured abstracts.

(a)

<table>
<thead>
<tr>
<th>Model</th>
<th>Complete</th>
<th>Partial</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-HMM</td>
<td>.247</td>
<td>.371</td>
</tr>
<tr>
<td>HMM</td>
<td>.226</td>
<td>.314</td>
</tr>
<tr>
<td>HMM + LDA</td>
<td>.217</td>
<td>.279</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Section</th>
<th>Complete</th>
<th>Partial</th>
<th>McKnight and Srinivasan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>Prec</td>
<td>Rec</td>
</tr>
<tr>
<td>Introduction</td>
<td>.923</td>
<td>.739</td>
<td>.723</td>
</tr>
<tr>
<td>Methods</td>
<td>.905</td>
<td>.841</td>
<td>.793</td>
</tr>
<tr>
<td>Results</td>
<td>.899</td>
<td>.913</td>
<td>.857</td>
</tr>
<tr>
<td>Conclusions</td>
<td>.911</td>
<td>.639</td>
<td>.847</td>
</tr>
</tbody>
</table>
significantly more training examples than our corpus of 27,075 abstracts, which could be a source of performance differences. Furthermore, details regarding their procedure for mapping structured abstract headings to one of the four general labels was not discussed in their paper. Nevertheless, our HMM-based approach is at least competitive with SVMs, perhaps better in some cases.

The results of our third set of experiments (training on RCTs and testing on a held-out testset of hand-annotated abstracts) is shown in Tables 3(a) and 3(b). Mirroring the presentation format above, Table 3(a) shows the classification error for the four-way label assignment problem. We noticed that some unstructured abstracts are qualitatively different from structured abstracts in that some sections are missing. For example, some unstructured abstracts lack an introduction, and instead dive straight into methods; other unstructured abstracts lack a conclusion. As a result, classification error is higher in this experiment than in the cross-validation experiments. We report performance figures for 35 abstracts that contained all four sections (“complete”) and for 14 abstracts that had one or more missing sections (“partial”). Table 3(b) reports accuracy, precision, recall, and F-measure for four separate binary classifiers (HMM with LDA) specifically trained for each section (one per row in the table). The table also presents the closest comparable experimental results reported by M&S—over 206 hand-annotated unstructured abstracts. Interestingly, M&S did not specifically note missing sections in their testset.

4 Discussion

An interesting aspect of our generative approach is that we model HMM outputs as Gaussian vectors (log probabilities of observing entire sentences based on our language models), as opposed to sequences of terms, as done in (Barzilay and Lee, 2004). This technique provides two important advantages. First, Gaussian modeling adds an extra degree of freedom during training, by capturing second-order statistics. This is not possible when modeling word sequences, where only the probability of a sentence is actually used in the HMM training. Second, using continuous distributions allows us to leverage a variety of tools (e.g., LDA) that have been shown to be successful in other fields, such as speech recognition (Evermann et al., 2004).

Table 2(b) represents the closest head-to-head comparison between our generative approach (HMM with LDA) and state-of-the-art results reported by M&S using SVMs. In some ways, the results reported by M&S have an advantage because they use significantly more training examples. Yet, we can see that generative techniques for the modeling of content structure are at least competitive—we even outperform SVMs on detecting “methods” and “results”. Moreover, the fact that the training and testing of HMMs have linear complexity (as opposed to the quadratic complexity of SVMs) makes our approach a very attractive alternative, given the amount of training data that is available for such experiments.

Although exploration of the tradeoffs between generative and discriminative machine learning techniques is one of the aims of this work, our ultimate goal, however, is to build clinical systems that provide timely access to information essential to the patient treatment process. In truth, our cross-validation experiments do not correspond to any meaningful naturally-occurring task—structured abstracts are, after all, already appropriately labeled. The true utility of content models is to structure abstracts that have no structure to begin with. Thus, our exploratory experiments in applying content models trained with structured RCTs on unstructured RCTs is a closer approximation of an extrinsically-valid measure of performance. Such a component would serve as the first stage of a clinical question answering system (Demner-Fushman and Lin, 2005) or summarization system (McKeown et al., 2003). We chose to focus on randomized controlled trials because they represent the standard benchmark by which all other clinical studies are measured.

Table 3(b) shows the effectiveness of our trained content models on abstracts that had no explicit structure to begin with. We can see that although classification accuracy is lower than that from our cross-validation experiments, performance is quite respectable. Thus, our hypothesis that unstructured abstracts are not qualitatively different from structured abstracts appears to be mostly valid.
5 Related Work

Although not the first to employ a generative approach to directly model content, the seminal work of Barzilay and Lee (2004) is a noteworthy point of reference and comparison. However, our study differs in several important respects. Barzilay and Lee employed an unsupervised approach to building topic sequence models for the newswire text genre using clustering techniques. In contrast, because the discourse structure of medical abstracts is well-defined and training data is relatively easy to obtain, we were able to apply a supervised approach. Whereas Barzilay and Lee evaluated their work in the context of document summarization, the four-part structure of medical abstracts allows us to conduct meaningful intrinsic evaluations and focus on the sentence classification task. Nevertheless, their work bolsters our claims regarding the usefulness of generative models in extrinsic tasks, which we do not describe here.

Although this study falls under the general topic of discourse modeling, our work differs from previous attempts to characterize text in terms of domain-independent rhetorical elements (McKeown, 1985; Marcu and Echihabi, 2002). Our task is closer to the work of Teufel and Moens (2000), who looked at the problem of intellectual attribution in scientific texts.

6 Conclusion

We believe that there are two contributions as a result of our work. From the perspective of machine learning, the assignment of sequentially-occurring labels represents an underexplored problem with respect to the generative vs. discriminative debate—previous work has mostly focused on stateless classification tasks. This paper demonstrates that Hidden Markov Models are capable of capturing discourse transitions from section to section, and are at least competitive with Support Vector Machines from a purely performance point of view.

The other contribution of our work is that it contributes to building advanced clinical information systems. From an application point of view, the ability to assign structure to otherwise unstructured text represents a key capability that may assist in question answering, document summarization, and other natural language processing applications.

Much research in computational linguistics has focused on corpora comprised of newswire articles. We would like to point out that clinical texts provide another attractive genre in which to conduct experiments. Such texts are easy to acquire, and the availability of domain ontologies provides new opportunities for knowledge-rich approaches to shine. Although we have only experimented with lexical features in this study, the door is wide open for follow-on studies based on semantic features.

7 Acknowledgments

The first author would like to thank Esther and Kiri for their loving support.

References


