Section Classification in Clinical Notes using Supervised Hidden Markov Model

Ying Li, Sharon Lipsky Gorman, Noémie Elhadad
Department of Biomedical Informatics
Columbia University
New York, NY 10032
{yil7003,srg7002,noemie}@dbmi.columbia.edu

ABSTRACT

As more and more information is available in the Electronic Health Record in the form of free-text narrative, there is a need for automated tools, which can process and understand such texts. One first step towards the automated processing of clinical texts is to determine the document-level structure of a patient note, i.e., identifying the different sections and mapping them to known section types automatically. This paper considers section mapping as a sequence-labeling problem to 15 possible known section types. Our method relies on a Hidden Markov Model (HMM) trained on a corpus of 9,679 clinical notes from NewYork-Presbyterian Hospital. We compare our method to a state-of-the-art baseline, which ignores the sequential aspect of the sections and considers each section independently of the others in a note. Experiments show that our method outperforms the baseline significantly, yielding 93% accuracy in identifying sections individually and 70% accuracy in identifying all the sections in a note, compared to 70% and 19% for the baseline method respectively.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Experimentation

Keywords

Clinical Natural Language Processing, Discourse Analysis, Section Labeling, Hidden Markov Model

1. INTRODUCTION

As more and more information is available in the Electronic Health Record in the form of free-text narrative, there is a need for automated tools, which can process and understand such texts. Equipped with such tools, data mining algorithms can be developped across multiple patient records to discover population-based trends, clinical decision support systems can incorporate the information entered in the patient notes by physicians into their reasoning module [3], and patient safety can be improved by the acurate recognition of a patient's condition [7].

There are several challenges entailed in machine understanding of clinical notes. Each one affects different levels of linguistic processing: determining the layout of a note (paragraphs, lists, and tables), identifying its discourse organization, syntactic parsing, mapping words or groups of words to semantic concepts in biomedical ontologies, and recognizing accurate modifiers and relations among concepts (co-references and temporal relations). Most of the research in clinical NLP so far has been carried out on the semantic parsing of patient notes.

This paper focuses on the discourse-level analysis of clinical notes, namely argumentative zoning [19]. Given a note composed of delimited sections, our system identifies its discourse organization by classifying each section according to one of several known section types. While researchers have investigated argumentative zoning in the context of scientific articles [20, 21], biomedical abstracts [9, 13, 14, 8, 18, 16, 6, 10] and news stories [1], little work has been conducted on discourse analysis of clinical notes besides the work of Denny and colleagues [4].

The accurate identification of the discourse structure of a clinical note can benefit several automated tasks, from word sense disambiguation and medication identification to data mining and text summarization. Word sense disambiguation in clinical notes is a challenging task. Generally in the English domain the assumption that one word has the same meaning throughout a document holds, but it is not the case in clinical notes. Within a note, for example, the acronym BS probably signifies "blood sugar" in the laboratory test section, but more likely signifies "breath sounds" in the physical examination section. Similarly, accurate tagging of medication information is influenced by the section in which it appears, as it is formatted differently depending on the section in which it is mentionned [5]. Wang and colleagues found that contextual filters based on sections had a positive impact on their accuracy of mining adverse drug events [23]. Finally, for automated problem list generation [11, 12] and more generally patient record summarization [22], it is essential to take into account the context of a term. For instance, information mentioned in the Family

IHI'10, November 11–12, 2010, Arlington, Virginia, USA.
Copyright 2010 ACM 978-1-4503-0030-8/10/11 ...$10.00.

CC: found down

HPI: 87W w/ HTN, osteoporosis, ? h/o falls with hip frx and bl THR found down by neighbor.

ALLERGIES: nkda

MEDICATIONS: ASA 81mg.

PMH:
- HTN
- Osteoporosis
- Hypercalcemia
- No abnormal results in 2005. Pt. does not want repeat.

PSHx:
- s/p L THR 8/99, r THR 2000
- s/p ORIF fx femur 9/2001
- Hx of PE in Rehab post ORIF 12/2001
- Colonic polyps
- R/L cataract extraction 2004
- R hip fx greater trachanter 4/15/2005.

Figure 1: Dummy extract of primary provider note.

History section of a note should be ignored when selecting patient-specific information.

Figure 1 shows an extract of primary provider note. In the interest of protecting patient privacy we show a mock-up of a real note. Typically, clinical notes contain sections about the following information: chief compliant, history of present illness, allergy, medication, past medical history, past surgical history, family history, social history, health care maintenance, review of system, physical exam, studies, laboratory tests and assessment & plan.

There are two main challenges in section classification for clinical notes. First, terms that physicians use to designate sections are ambiguous and various, for example, “history of present illness” might appear as “HPI,” “history” or “history of current illness.” Second, physicians often omit section headers when they author clinical notes (in our dataset, approximately two thirds of the notes do not have any headers). For both these reasons, a section classification system must be able to infer a section type given the text in the section. In this paper we hypothesize that knowledge of the ordering of the sections can improve the accuracy of a section classifier. To test this hypothesis, we train a Hidden Markov model (HMM) that categorizes sections in clinical notes into one of 15 pre-defined section labels.

The key contributions of this work are three-fold: (1) our method handles a large number of section labels compared to previous work in the biomedical literature; (2) it relies on simple lexical features encoded through language models; and (3) it relies on a sequence-based classifier, thus departing from previous work in clinical note discourse analysis which classifies each section independently of the other.

We show that our system outperforms a baseline system, which considers each section in isolation, in two evaluation setups: per-section accuracy (whether the system predicted the gold-standard label of a given section) and per-note accuracy (whether the system correctly predicted all the section labels within a note).

2. RELATED WORK

The task of argumentative zoning consists of classifying sentences in a text according to mutually exclusive categories. The categories convey discourse-level information and follow a rhetorically based schema [19]. The definition of such schema should be independent of the specific content of a text [21]. For instance, in a scientific article, a valid argumentative zoning schema can include zones such as background (B), objective (O), method (M), result (R), conclusion (C) and introduction (I). This schema is stable across articles aside from minor variations of labeling.

Previous work on identifying section labels include Naive Bayesian Models (NBM) [16], Support Vector Machine (SVM) [9], Hidden Markov Model (HMM) [8], Maximum Entropy (ME) [10], and Conditional Random Field (CRF) [6]. All studies above target abstracts in research papers and the classification operates at the sentence level. Ruch and colleagues classified biomedical abstract sentences into four section types (O, M, R, C) using linearly combined features. Features were stemmed unigrams, bigrams and trigrams [16]. McKnight and Srinivasan labeled abstract sentences of randomized clinical trials into four section types (I, M, R, C) through an SVM classifier. Features were bag of words and relative sentence location in the abstract [9]. Lin and colleagues annotated sentences with four labels (I, M, R, C) through a generative model. Features were a bigram language model for each of the four sections. The language models were smoothed with Kneser-Ney discounting and Katz backoff [8]. Merity and colleagues implement a maximum entropy model to label sentences according to seven categories of Teufel’s work. Features were n-grams, the first four words of a sentence, section counter, sentence position between two sections, sentence position within a paragraph, length and named entity information [10]. Finally, Hirohata and colleagues assigned four section labels (O, M, R, C) into abstract sentences. Features included n-gram including stemmed unigrams, bigrams and mixture of them, relative sentence location and features form previous / next n sentences [6].

In the clinical domain, the sole relevant work is the one of Denny and colleagues [4]. A Naive Bayes classifier is trained on a set of clinical notes with a large set of section labels. Our work contrasts from their approach in the following ways: (i) sections are classified as part of a sequence, not independently of the other sections in the note; (ii) our set of section labels is more generic than theirs, and thus smaller, so as to be robust across note types (e.g., discharge summaries vs. outpatient consult notes); and (iii) like in their work, our dataset is comprised of sections with headers mapped to the 15 section labels automatically according to a hand-built mapping dictionary. This allows us to rely on a large dataset of notes annotated with labels. However, we make sure when training and testing our methods to ignore the headers, so as to not influence the classifiers.

Besides the choice of model learning strategy, choice of features is important for the classification task. Argumentative zoning uses several traditional text classification features such as n-grams. Some researchers emphasized the importance of domain-independent features, while others combine a set of manually crafted expression, such as “we aim at,” “In conclusion,” and “our goal,” which are most common in Objective and Conclusion sentences in the scientific literature [20]. Mullen and Mizuta provided a baseline feature

Table 1: Most common section orderings in the corpus.

Rank	Nb. of Notes	Freq.	Section Sequence
1	646	6.7%	HPI → ALL → MEDS → PE → LABS → A/P
2	348	3.6%	PE → LABS → A/P
3	283	2.9%	HPI → ALL → PE → LABS → A/P
4	243	2.5%	MEDS → PE → A/P
5	217	2.2%	PMH → MEDS → A/P
6	179	1.8%	PE → A/P
7	171	1.8%	CC → HPI → ALL → MEDS → SHX → FHX → ROS → PE → LABS → A/P
8	110	1.1%	ALL → MEDS → SHX → FHX → ROS → PE → LABS → A/P
9	107	1.1%	PMH → ALL → MEDS → PE → A/P
10	106	1.1%	ALL → MEDS → PE → A/P
...	...	...	...
911	3	0.03%	MEDS → ALL → PMH → FHX → SHX → ROS → PE

set for learning rhetorical zones and evaluated the effective of features [13, 14]. The baseline feature set included lexical/syntactic information composed of features representing word unigrams, stemmed unigrams and bigrams, and dependency triples derived from a syntactic parser, main verb, location composed of the name of the section in which the sentence occurs and the absolute location, and zone sequence adding all of the above feature information for previous and subsequent sentences. The last set of features in effect implements a proxy for a sequential model.

Discriminative approaches have been shown to be very effective for supervised section classification task. However, their high computational complexity related to training size prohibits them for huge training set. In addition, generative model can be as the basic step for further application such as extractive summarization [1]. Barzilay and Lee's paper is the first work to employ a generative approach to directly model content and classify sentence into a specific topic [1]. Naive Bayes, SVM and maximum-entropy models consider the task of labeling section names as a text categorization that determines section label s_k for each text unit t_i. Each text unit is considered independently of the other units. In contrast, CRF and HMM formalize the section classification task as a sequential labeling problem.

3. METHODS

In this section we present our algorithm for assigning one of 15 section types. The section types are chief complaint (**CC**), assessment and plan (**A/P**), allergies (**ALL**), family history (**FHX**), social history (**SHX**), past medical history (**PMH**), past surgical history (**PSH**), past medical history and past surgical history (**P/P**), history of present illness (**HPI**), laboratory tests (**LABS**), physical examination (**PE**), review of system (**ROS**), studies (**STUDY**), medication (**MEDS**), and health care maintenance (**H/M**).

3.1 Section Identification as a Sequence Labeling Task

Like in abstracts of scientific articles, where position of a sentence influences the likelihood of a particular section type (the first sentence of an abstract, for instance, is likely to be classified as Objective, while the last belongs to the Conclusion type), clinical notes exhibit regularities in the order of sections they contain. However, because there is a large number of section types, and not all section types are guaranteed to be present in a particular note, there is a large variation of section orderings. To understand this point further, we enumerated all the section orderings occurring in our dataset of 9,679 clinical notes. The top 10 most common section orderings are listed along with one of the least common ones in Table 1. Out of about 9,000 notes, there are only 911 unique orderings, which confirms the intuition the presence of common patterns of sequence orderings. The most frequent sequence is "HPI → ALL → MEDS → PE → LABS → A/P," which is consistent with what physicians consider the typical order of information in a clinical note.

We would like to model the sequence of section labels and their textual content in clinical notes rather than considering each section independently of the others in the note. Thus, we formalize our problem as a sequence labeling task: given a clinical note with n sections $x = (x_1, ..., x_n)$, determine the optimal sequence of section labels $s^* = (s^*_1, ..., s^*_n)$ among all possible section sequences. Hidden Markov Models (HMM) have been very successful for a very large number of applications in natural language processing, biomedical informatics and numerous other fields. We describe next our HMM-based section classifier and specify the parameters of HMM based on Rabiner's tutorial [15] and Barzilay and Lee's work [1].

3.2 Hidden Markov Model Parameters

We introduce the terms *token* and *text span* when referring to the content of a clinical note. A token is defined as text separated by white space, while a text span consists of tokens which extend between two section labels. In Figure 1, section labels are indicated in boldface; "h/o," "hip," and "Pt." are tokens; "nkda" is both a token and a text span; "found down" is a text span. Moreover, ("CC," "HPI," "ALL," "MEDS," "PMH," "PSH") constitutes a particular section label ordering for that note. As such, a clinical note is represented as a sequence of text spans, each presumed to convey information about a specific section.

We train an HMM with 15 states, each state s corresponding in intuition to a distinct section label. For a given text span, the observations for each state are modeled according to a bigram language model specific to each section type. We depart from previous work in argumentative zoning, as we operate at the section level rather than sentence level. The state transition probabilities capture constraints on section orderings. We learn the observation and transition probabilities from a corpus of clinical notes annotated with section labels.

Observation Probabilities Computation. We train a language model for each section state s of the HMM. The language model is smoothed with Laplace smoothing. The likelihood of a n-token text span $x = t_1 t_2 ... t_n$ generated by a state s, is computed according to Equation 1. The specific state bigram probability can be estimated through Equation 2, which implements Laplace smoothing counts for the corresponding section s (V_s represents the vocabulary size for that section state). Smoothing is critical in our experiment since 23% of the bigram tokens ($t_k t_{k+1}$) in the test set are unseen in the training set [2]. All tokens in the training set are lowercase; all numbers are converted into a generic symbol "NUM." We did not implement common preprocessing techniques, such as normalizing tokens or excluding stopwords.

$$P_s(x) = \prod_{i=1}^{n} P_s(t_{i+1}|t_i) \tag{1}$$

$$P_s(t_{i+1}|t_i) = \frac{count_s(t_i t_{i+1}) + 1}{count_s(t_i) + |V_s|} \tag{2}$$

Transition Probabilities Computation. The state transition probabilities are estimated by Equation 3, which is smoothed by the total number of section labels.

$$P(s_j|s_i) = \frac{count(s_i s_j) + 1}{count(s_i) + 15} \tag{3}$$

Decoding. In order to identify the optimal section label sequence as defined in Equation 4 associated with the given observation sequence, we rely on the Viterbi algorithm.

$$(s_1^* ... s_T^*) = argmax_{1 \leq i_1, i_T \leq 15} P(s_{i_1} ... s_{i_T} | x_1 ... x_T, model) \tag{4}$$

3.3 Baseline System

We hypothesize that section sequences can improve the accuracy of section label classification. In order to test this hypothesis, we introduce the following baseline. Each section is classified independently of the other sections in the note. Given a text span x, the selected section s^* is the one which yields the largest observation probability (Equations 1 and 5).

$$s^* = argmax_s P_s(x) \tag{5}$$

4. EXPERIMENTS

4.1 Corpus and Data Preparation

In classification tasks, there are two competing strategies – expert-driven and data-driven approaches. The former relies on a domain expert, which is both time and labor intensive, and the latter directly depends on the availability of large training sets. Fortunately, labeling for training and testing purposes can be acquired in a cheap fashion. Our corpus consists of outpatient clinical notes for years 2008 to 2009 from the data repository of the NewYork-Presbyterian Hospital. These include many note types, including primary provider notes, consultation notes, follow up, and clinical summaries. Most of the notes in corpus were unstructured (i.e., they do not have any explicit section headers). In fact, the ratio of notes with sections headers and without headers is 33% vs. 67%. Our dataset of 9,679 notes contains notes with section headers, so we can have a fully annotated dataset.

There are two steps to extract section headers and their respective sections. First, text span boundaries are determined by using section headers and blank lines as textual cues. A text span may start with a section header or it may start with the occurrence of a blank line. The start position of next text span becomes the marker of the end of the last text span. Second, a small collection of four pattern matching rules are applied to recognize candidate section labels after the detection of text span boundary. Finally, recognized section headers are mapped to section labels based on a predefined section label dictionary.

The mapping dictionary has on average more than 10 possible lexicalizations for each section label. For instance, the headers "Treatment plan," "impression/plan," and "assessment and plan" all mapped to "A/P;" "soc h," "sochx," and "socialh" mapped to "SHX;" and "surgeries," "psurghx," and "pshx" mapped to "PSH." In order to evaluate the accuracy of the dictionary-based mapping, two physicians were asked to annotate 120 clinical notes randomly selected from our corpus. Overall, there were 866 sections. The dictionary match achieves high accuracy of 97.36%. We interpret this result as a confirmation that our annotated training and test sets contain valid annotation, or at least annotation valid enough for learning purposes.

In our dataset of 9,679 notes we paid attention to the following: we did not include notes with less than two section headers; we did not include notes in the training data, which came from the same patient id as the clinical notes present in the test set. This way, we avoid boosting the accuracy of our classifier artificially due to repetition of content in the notes for the same patient. Finally, we split the corpus into training/test sets according to the appearance of section sequence. Clinical notes with the same section ordering were divided into training set and test data respectively. Overall, the training set comprised 78% of notes in the corpus (thus, the testing set is the remaining 22%).

4.2 Results

We now describe the evaluation of our section classifier on the test set. The test set contains 2,088 clinical notes, corresponding to 11,706 text spans. Thus, on average, each clinical note had 5.6 sections.

We compare our HMM-based classifier to the baseline system. We report precision, recall, F-measure for each section label (Table 2), the overall micro average values, the overall macro average values and accuracy (Table 3) as our evaluation metrics at the section level [17]. We also report on per-note accuracy (Table 3). In this case, a true-positive is a note with *all* its sections correctly labelled.

The HMM outperforms the baseline in most of metrics except the recall for CC labels. Table 4 shows the confu-

Table 2: Performance of the HMM and the Baseline classifiers (n=11,706).

	Precision		Recall		F-measure	
	HMM	Baseline	HMM	Baseline	HMM	Baseline
A/P	0.85	0.81	0.96	0.51	0.90	0.63
ALL	0.97	0.82	0.93	0.92	0.95	0.86
CC	0.96	0.18	0.57	0.69	0.72	0.28
FHX	0.98	0.67	0.95	0.89	0.97	0.76
SHX	0.98	0.92	0.98	0.90	0.98	0.91
H/M	0.90	0.47	0.91	0.91	0.91	0.62
HPI	0.76	0.44	0.85	0.34	0.80	0.38
LABS	0.97	0.89	0.95	0.74	0.96	0.81
MEDS	0.99	0.98	0.97	0.84	0.98	0.90
PE	0.99	0.97	0.99	0.82	0.99	0.89
PMH	0.92	0.76	0.87	0.55	0.89	0.64
P/P	0.72	0.11	0.68	0.37	0.70	0.17
PSH	0.94	0.74	0.89	0.77	0.91	0.76
ROS	0.93	0.62	0.86	0.79	0.89	0.70
STUDY	0.94	0.63	0.82	0.66	0.88	0.65

sion matrix for the HMM classification. Rows present the predicted labels and columns represent the gold-standard labels. The diagonal indicates the ratio of correctly classified sections for each of the section labels. Confusion matrices are helpul to identify cross-section confusions and conduct error analysis. We observe that STUDY sections are often misclassified into LABS (10.24%). We interpret this as a limitation in the sequential aspect of the data, as STUDY and LABS sections are very often adjacent in the notes. We further observe that CC sections are often misclassified into A/P (23.36%) and HPI (15.75%), while HPI sections are often misclassified into A/P (12.56%). At the same time, the classifier learns to annotate a text span as A/P, HPI and CC when it makes a mistake (i.e., A/P has high false positive). All of them make sense in the medical practice: the chief complaint section records the patient's subjective narrative of his/her physical condition; the history of present illness is about physician's observation and objective observation about the patient's physical condition; and the assessment & plan is about treatment protocols according to the patient's physical condition and physician's diagnosis. There is a strong content overlap among these three section types. Moreover, some tokens in other sections are similar to CC, HPI and A/P since these three sections are overviews of patients' physical conditions. It is obvious that sometimes the observation probabilities have stronger impact on the HMM than the transition probabilities, based on the evidence that the system often misclassifies among CC, HPI and A/P: In practice, A/P is at the end of the clinical notes, HPI and CC are often at the beginning of the clinical notes. Thus, the most probable previous states are different for these three states. Although transition matrix distinguishes them well, observation matrix counteracts the discrimination. A potential way to address this limitation is to experiment with different number of states. In our method, we build the HMM with 15 states, one state per section label, but the number of states is a parameter of the method.

Finally, both the micro F-measure and the macro F-measure are above 90%, statistically significantly above the baseline's, which are about 70% (Table 3) ($p < 0.0001$). Per-note accuracy reaches 70% in HMM compared with 19% for the baseline. It reflects the ability of HMMs to determine the optimal sequence of section labels. The micro-average measures take into account the prevalence of each section type in the test set. The fact that they are higher than the macro-average measures indicate that our method appropriately fits the characteristics of our dataset.

Table 3: Overall average performance of the HMM and the Baseline classifiers. Statistical significance at p<0.0001 (n=11,706 for per-section and n=2,088 for per-note).

	Baseline	HMM
Micro P	0.71	0.93*
Micro R	0.71	0.93*
Micro F	0.71	0.93*
Macro P	0.67	0.92*
Macro R	0.72	0.88*
Macro F	0.69	0.90*
Per Section Acc.	0.71	0.93*
Per Note Acc	0.19	0.70*

Overall, our HMM-based classifier provides an appropriate framework for section classification in clinical notes, as it captures the ordering constraints in clinical notes. Furthermore, section-dependent language models are appropriate models for deriving the observation probabilities. When examining the data set, the distinction between some section types appears questionable (for instance, PMH and P/P). Merging such sections could be both beneficial and legitimate. Interestingly, in the manual annotation of section headers carried out by the two physicians for the evaluation of our mapping dictionary, most of the disagreements between our mapping and theirs were also about PMH and P/P.

Finally, we have assumed that text span boundaries in clinical notes are known, which is unusual in practice. In order to improve the accuracy of the classifier, a future direction for this study is to segment and classify text spans at the same time.

Table 4: Confusion matrix. Rows are the predicted labels and columns are the true labels. Results are in percent.

	A/P	CC	HPI	ALL	FHX	SHX	H/M	LABS	MEDS	PE	PMH	P/P	PSH	ROS	STUDY
A/P	95.82	23.36	12.56	1.41	1.48	0.25	5.33	2.77	1.12	0.22	7.07	5.06	2.15	1.50	4.88
CC	0.09	57.22	0.12	0.40	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.31	0.00	0.00
HPI	2.96	15.75	85.31	1.21	0.82	0.74	0.00	0.23	0.37	0.22	2.14	5.06	2.15	10.03	1.46
ALL	0.00	0.26	0.00	92.94	0.16	0.00	0.00	0.12	0.50	0.11	0.11	0.00	0.00	0.25	0.00
FHX	0.00	0.00	0.00	0.00	95.57	0.12	0.00	0.12	0.06	0.06	0.21	0.00	0.62	0.25	0.00
SHX	0.14	0.00	0.00	0.81	0.16	98.53	0.00	0.00	0.19	0.00	0.11	0.00	0.31	0.25	0.00
H/M	0.23	0.52	0.24	0.20	0.00	0.00	91.11	0.46	0.19	0.06	0.11	1.27	0.00	0.25	0.00
LABS	0.00	0.00	0.00	0.00	0.00	0.00	0.44	95.72	0.00	0.11	0.00	0.00	0.00	0.00	10.24
MEDS	0.09	0.00	0.00	1.41	0.33	0.00	0.00	0.00	97.01	0.06	0.21	0.00	0.62	0.25	0.00
PE	0.00	0.26	0.47	0.20	0.00	0.12	0.89	0.00	0.00	98.94	0.21	0.00	0.31	0.75	0.00
PMH	0.42	1.31	0.59	0.60	0.33	0.00	2.22	0.23	0.25	0.00	86.60	16.46	4.00	0.25	0.98
P/P	0.00	0.00	0.36	0.00	0.33	0.00	0.00	0.00	0.00	0.00	1.50	68.35	0.31	0.25	0.00
PSH	0.00	0.26	0.12	0.60	0.16	0.12	0.00	0.00	0.19	0.00	0.75	2.53	88.92	0.00	0.00
ROS	0.05	1.05	0.12	0.20	0.66	0.12	0.00	0.00	0.00	0.22	0.75	1.27	0.31	85.96	0.00
STUDY	0.19	0.00	0.12	0.00	0.00	0.00	0.00	0.35	0.00	0.00	0.21	0.00	0.00	0.00	82.44

5. CONCLUSION

To our knowledge, this paper is the first to introduce a generative sequence-based model to identify section labels in clinical notes. The proposed method outperforms a baseline bigram model, which ignores sequence information. The HMM has roughly 20% point increase in F-measure for per-section accuracy and 50% point increase in per-note accuracy over the baseline. Our features are simple: the bigram language model of text spans for each state of the HMM.

6. ACKNOWLEDGMENTS

This work was supported by a National Library of Medicine grant R01 LM010027 (NE). Any opinions, findings, or conclusions are those of the authors, and do not necessarily reflect the views of the funding organization.

7. REFERENCES

[1] R. Barzilay and L. Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of HLT-NAACL*, pages 113–120, 2004.

[2] S. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4):359–394, 1999.

[3] D. Demner-Fushman, W. Chapman, and C. McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, 2009.

[4] J. Denny, A. Spickard, K. Johnson, N. Peterson, J. Peterson, and R. Miller. Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association*, 16(6):806, 2009.

[5] S. Gold, N. Elhadad, X. Zhu, J. Cimino, and G. Hripcsak. Extracting structured medication event information from discharge summaries. In *Proceedings of the AMIA Annual Symposium*, pages 237–241.

[6] K. Hirohata, N. Okazaki, S. Ananiadou, M. Ishizuka, and M. Biocentre. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of IJCNLP*, pages 381–388, 2008.

[7] G. Hripcsak, S. Bakken, P. Stetson, and V. Patel. Mining complex clinical data for patient safety research: a framework for event discovery. *Journal of Biomedical Informatics*, 36(1-2):120–130, 2003.

[8] J. Lin, D. Karakos, D. Demner-Fushman, and S. Khudanpur. Generative content models for structural analysis of medical abstracts. In *Proceedings of the HLT-NAACL BioNLP Workshop*, pages 65–72, 2006.

[9] L. McKnight and P. Srinivasan. Categorization of sentence types in medical abstracts. In *Proceedings of the AMIA Annual Symposium*, pages 440–444, 2003.

[10] S. Merity, T. Murphy, and J. Curran. Accurate argumentative zoning with maximum entropy models. In *Proceedings of the ACL Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 19–26, 2009.

[11] S. Meystre and P. Haug. Automation of a problem list using natural language processing. *BMC Medical Informatics and Decision Making*, 5(1):30, 2005.

[12] S. Meystre and P. Haug. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *Journal of Biomedical Informatics*, 39(6):589–599, 2006.

[13] Y. Mizuta and N. Collier. Zone identification in biology articles as a basis for information extraction. In *Proceedings of the COLING Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 29–35, 2004.

[14] T. Mullen, Y. Mizuta, and N. Collier. A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. In *ACM SIGKDD Explorations Newsletter*, volume 7, page 58, 2005.

[15] L. Rabiner. A tutorial on hidden markov models and selected applications inspeech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, 1989.

[16] P. Ruch, C. Boyer, C. Chichester, I. Tbahriti, A. Geissbuhler, P. Fabry, J. Gobeill, V. Pillet, D. Rebholz-Schuhmann, and C. Lovis. Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics*, 76(2-3):195–200, 2007.

[17] F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys*, 34(1):1–47, 2002.

[18] I. Tbahriti, C. Chichester, F. Lisacek, and P. Ruch. Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the medline digital library. *International Journal of Medical Informatics*, 75(6):488–495, 2006.

[19] S. Teufel, J. Carletta, and M. Moens. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of EACL*, pages 110–117, 1999.

[20] S. Teufel and M. Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

[21] S. Teufel, A. Siddharthan, and C. Batchelor. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings of EMNLP*, pages 1493–1502, 2009.

[22] T. V. Vleck, A. Wilcox, P. Stetson, S. Johnson, and N. Elhadad. Content and structure of clinical problem lists: A corpus analysis. In *Proceedings of the AMIA Annual Symposium*, pages 753–757.

[23] X. Wang, H. Chase, M. Markatou, G. Hripcsak, and C. Friedman. Selecting information in electronic health records for knowledge acquisition. *Journal of Biomedical Informatics*, 2010.