Modeling Physicians' Utterances to Explore Diagnostic Decision-making

Xuan Guo, Rui Li, Qi Yu, Anne R. Haake
Rochester Institute of Technology
{xxg3358, rxlics, qi.yu, arhics}@rit.edu

Abstract

Diagnostic error prevention is a long-established but specialized topic in clinical and psychological research. In this paper, we contribute to the field by exploring diagnostic decision-making via modeling physicians' utterances of medical concepts during image-based diagnoses. We conduct experiments to collect verbal narratives from dermatologists while they are examining and describing dermatology images towards diagnoses. We propose a hierarchical probabilistic framework to learn domain-specific patterns from the medical concepts in these narratives. The discovered patterns match the diagnostic units of thought identified by domain experts. These meaningful patterns uncover physicians' diagnostic decision-making processes while parsing the image content. Our evaluation shows that these patterns provide key information to classify narratives by diagnostic correctness levels.

1 Introduction

Studies in the field of psychology and health sciences show evidence that overconfidence may cause diagnostic error when a physician overlooks key clues and prematurely reaches a diagnosis [Berner and Graber, 2008]. This can be explained by a case in the *dual-process theory* where the physician's *intuitive* system dominates her decision-making process over *analytical* system [Croskerry, 2009]. To disentangle the underlying factors that may relate to diagnostic correctness and diagnostic confidence, we collect physicians' verbal narratives during image-based diagnoses and model the uttered medical concepts for their reasoning processes.

We first design two experiments to elicit physicians' diagnostic verbal narratives which contain their verbal descriptions of image content when inspecting each photographic dermatological image toward a diagnosis (Section 3.1). Experiment I contains additional labels for a diagnostic correctness study, and Experiment II for a diagnostic confidence study (Section 3.3). To extract meaningful behavioral patterns, we then model physicians' use of domain concepts and their reasoning order of these concepts during diagnoses. In particular, we remove non-medical tokens in both studies using the unified medical language system (UMLS) [Fung and Bodenreider, 2012], so as to only consider physicians' utterances of medical concepts in each verbal narrative (Section 3.2). The remaining medical concept sequences in the two studies are separately modeled using a hierarchical probabilistic framework we develop (Section 4). In this manner we discover the stereotypical and idiosyncratic patterns that commonly exist in both studies, and these patterns represent the verbal narratives at a higher level. In this automatically discovered high-level representation, the diagnostic decision-making processes are quantified and visualized (Section 5).

To evaluate and interpret the patterns in the application domain, we compare them with the *diagnostic thought units* defined and identified by medical doctors [McCoy *et al.*, 2012] (Section 5.1). These thought units are standardized semantic labels to abstract the medical terms per their positions and functions in diagnostic reasoning processes. Table 1 lists some examples. We also use the additional labels collected from physicians during or after experiments to identify groups of diagnostic narratives in terms of the diagnostic correctness and confidence (Section 3.3). An evaluation study shows the importance of the discovered patterns to classify narratives into different correctness levels (Sections 5.2). This paper contributes to research in the field as follows:

- We naturally elicit expert spoken narratives through in-scenario experiments.
- We extract domain-specific concepts from narratives to allow modeling at the semantic level.
- We model expert' utterances of medical concepts during image-based diagnoses to gain insights into cognitive reasoning strategies.

	Thought Unit Labels (Abbr.)	Instances
Type I	Patient DEMographics (DEM)	elderly, caucasian, woman
	Body LOCation (LOC)	arm, upper lip, knuckles
	Lesion CONfiguration (CON)	linear, annular, grouped
	SECondary finding (SEC)	crust, ulcer, erythematous
	Lesion DIStribution (DIS)	solitary, bilateral, extensive
Type II	PRImary lesion type (PRI)	papule, plaque, patch
	DIFferential diagnosis (DIF)	X, Y or Z
	Final Diagnosis (DX)	this is X

Table 1: Two types of thought units.

2 Related Work

Current research on diagnostic accuracy relies on the analysis of either research interviews or clinical chart records [Galanter and Patel, 2005]. For example, Bowen studied reports of clinicians and medical students to explore educational strategies to transfer classroom knowledge to clinical decision-making [Bowen, 2006]. We propose an objective data collection paradigm and an automated approach to analyzing diagnostic decision-making.

Natural language processing models, such as bag-of-word and N-gram, have been used to analyze clinical texts. Since different practitioners express similar meanings in a variety of ways both syntactically and lexically, the medical datasets tend to be sparse. To tackle this, topic modeling approaches, such as latent semantic analysis (LSA) [Deerwester *et al.*, 1990] and latent Dirichlet allocation (LDA) [Blei *et al.*, 2003], transform original vocabulary to latent variables (a.k.a., topics) whose mixtures summarize the documents more abstractly. Despite the advantage to produce high-level representations, topic modeling approaches do not consider the word order within each document.

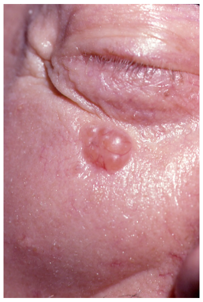
To recognize temporal patterns in the documents and speech data, hidden Markov model (HMM) can be used [Rabiner, 1989]. It learns a sequential structure of hidden states (i.e., patterns), each being a probability distribution over the vocabulary. We propose to use HMMs to model physicians' spoken narratives, because the order of thoughts (i.e., cognitive states) is crucial to diagnostic decision-making [Croskerry, 2009; Berner and Graber, 2008]. To automatically determine the optimal number of hidden states, Teh et al. developed a Bayesian non-parametric HMM using hierarchical Dirichlet processes [Teh *et al.*, 2006] and Van Gael et al. developed the *beam sampler* for it to limit the computational costs [Van Gael *et al.*, 2008]. This paper extends Teh et al.'s model and the beam sampler to a multi-sequence variant to allow learning from a group of medical concept sequences (see Section 4). This results in a desired sequential representation, based on which we train classifiers to differentiate narrative groups.

3 Datasets

3.1 Data Elicitation Experiments

To obtain expert data, we conduct two data elicitation experiments where dermatology images are presented as visual stimuli for physicians to inspect. During the experiments, each physician is instructed to describe the image content to a student seated nearby as if teaching.

- Experiment I contains 48 dermatology images covering a wide range of diagnoses, and it is used to explore diagnostic decision-making that relate to diagnostic correctness. We record 16 physicians' verbal image descriptions in this experiment, and hence we have $48 \times 16 = 768$ verbal description trials.
- Experiment II focuses on only a few disease categories by 30 images, and it is used to study the decision making related to diagnostic confidence. There are 29 physicians, and after removing 3 trials due to data collection failures we have 867 verbal description trials.



(a) Basal cell carcinoma diagnostic image case (image courtesy of Dr. Cara Calvelli).

(SIL) um this is a (SIL) pearly lobulated (SIL) pink (SIL) papule with telangiectasias multi-lobulated papule with telangiectasias (SIL) on the upper cheek (SIL) of (SIL) an elderly individual (SIL) say nodular basal cell carcinoma it's a nodular basal cell carcinoma (SIL) uh (SIL) there is some surrounding sun damage (SIL)

(b) A dermatologist's narrative for the image. *(SIL)* represents silent pause.

PRI ; SEC ; LOC ; DEM ; DX or DIF

Figure 1: A sample image (a) and a corresponding diagnostic narrative (b) annotated by thought unit labels.

3.2 Linguistic Data Preprocessing

All verbal description trials are transcribed as diagnostic verbal narratives with tokens and time-stamps included using the speech analysis tool Praat [Boersma and Weenink, 2009]. We use a medical knowledge source (i.e., the UMLS) and a Metathesaurus mapping tool (i.e., the MetaMap [Aronson, 2001]) to remove non-medical words and join adjacent words (e.g., *basal*, *cell*, and *carcinoma*) into multiwords (e.g., *basal cell carcinoma*) when necessary (see Figure 1-b). This turns each narrative to a sequence of medical concepts.

3.3 Gold Standard

Thought units: McCoy et al. collected medical doctors' annotations that partition and label diagnostic reasoning records into meaningful units of thought [McCoy *et al.*, 2012]. These thought units cover the terminology to standardize the description of skin lesions, including lesion arrangement, distribution, texture, color, primary lesion type, and diagnosis [Lyons and Ousley, 2014]. We use this thought unit labeling as a gold standard in our study to evaluate and interpret the patterns discovered by the model. Since our image set contains a wider range of diagnoses, their thought unit labels do not cover our whole vocabulary.

Diagnostic correctness levels: We recruit three dermatologists to evaluate the narratives from the 16 participating physicians in Experiment I in terms of their diagnostic correctness. A correctness score is assigned to each narrative, which balances the correctness of described Type II thoughts (i.e., primary lesion type, differential diagnosis, and final diagnosis). This score ranges from 0 to 3 and its distribution across narratives is in Figure 2. We define the correctness score below 1 (inclusive) as low-correctness and that above 2 (inclusive) high-correctness. These two levels of narratives in Experiment I are classified using the patterns discovered by the model, and the differences in narration patterns between two classes are visualized in Section 5.

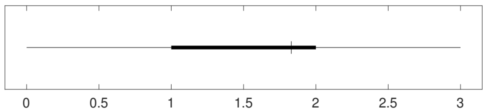


Figure 2: The correctness score distribution across all narratives.

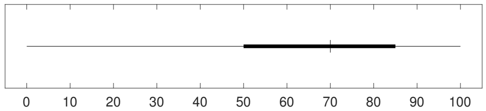


Figure 3: The self-reported diagnostic confidence score distribution across all narratives.

Diagnostic confidence levels: During Experiment II, each participating physician is required to report her diagnostic confidence at the end of narration. Figure 3 shows how the diagnostic confidence scores are distributed across all narratives. We define the bottom quartile (0%–50% confidence, inclusive) as low confidence and the top quartile (85%–100% confidence, inclusive) high confidence.

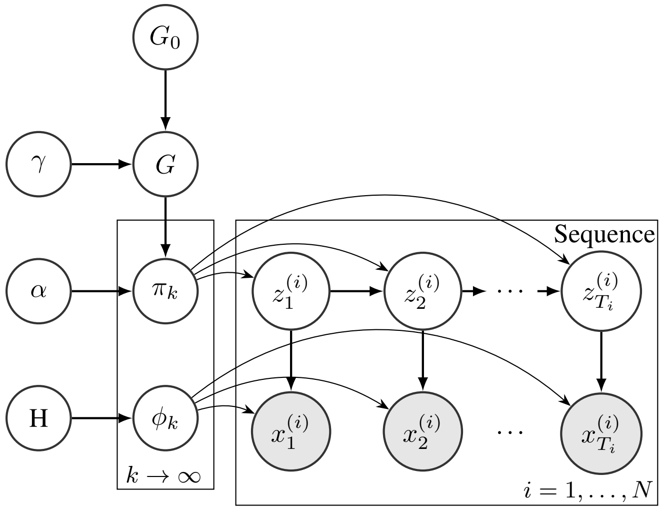


Figure 4: The hierarchical Dirichlet process-hidden Markov model that learns from multiple medical concept sequences as a group.

4 Hierarchical Dynamical Model

4.1 Hierarchical Dirichlet Process

LDA can be used to discover latent topics for document modeling, where each *topic* is a distribution of terms in vocabulary. For each document, a mixture of topics is drawn from a Dirichlet distribution, and then each term in the document is drawn independently from that mixture. The hierarchical Dirichlet processes (HDP) mixture model is a nonparametric generalization of LDA to automatically determine the number of mixture components (topics), which is not known a priori [Teh *et al.*, 2006]. This enables the number of topics to be unbounded. HDP uses a Dirichlet process to capture the uncertainty in the number of topics.

In particular, the Dirichlet process DP(α, G), specified by a base distribution G and a concentration parameter α, characterizes how the random variables are distributed according to G. The base distribution G is the expected value of the process, and it is selected to represent the countably-infinite set of possible topics for the corpus, and then the finite distribution of topics for each document is sampled from this base distribution. The Dirichlet process models data that tend to repeat previous values in a *rich get richer* manner. More specifically, each expected value is generated in proportion to $\frac{n_i}{\alpha+\sum_i n_i}$ (n denotes the number of previous occurrences and i indexes each expected value). The concentration parameter α is a positive real number that indicates the probability proportion to generate a new value, $\frac{\alpha}{\alpha+\sum_i n_i}$, which enables the Dirichlet process to model variables of unknown cardinality.

The DP models a group of data by variables of unknown cardinality, and the HDP is useful to address the problems that have multiple groups of data by tying the variables across groups [Teh *et al.*, 2006]. Each group of data is modeled using a mixture model, with mixture components shared across all groups but mixing proportions being group-specific. The basic building block of hierarchical Dirichlet processes is recursion in which the base measure G for a Dirichlet process $G_j \sim$ DP(α, G) is itself a draw from a Dirichlet process $G \sim$ DP(γ, G_0). This recursive construction enforces the random measure G_j to place its atoms at the discrete locations determined by G.

4.2 HDP-HMM

To address the sequential property of the data, a canonical HMM specifies a set of finite mixture distributions, one for each value of the current state z_t. Given z_t, the observation x_{t+1} is chosen by first generating the state z_{t+1} and then generating x_{t+1} conditional on z_{t+1}. A Dirichlet process can be used to replace the generating process of the finite mixture model. However, in order to tie up all the potential states to be able to access one from another, the hierarchical Dirichlet processes has to be used [Teh *et al.*, 2006]. This forms the HDP-HMM, or infinite HMM (iHMM).

4.3 M-seq HDP-HMM

We develop a HDP-HMM variant that incorporates observations of multiple sequences, and we call it *M-seq HDP-HMM*, or *M-seq iHMM*.

Particularly, since the preprocessing does not affect the sequential order of the remaining medical concepts in the narratives, we use HMMs as the likelihood to characterize the temporal dynamic nature of the medical concept sequences. In Figure 4, each learned hidden state sequence $\{z_t^{(i)}\}_{t=1,2,\ldots,T_i}$ presents a subset of all the hidden states that particularly corresponds to the observed medical concept sequence $\{x_t^{(i)}\}_{t=1,2,\ldots,T_i}$. We use the hierarchical Dirichlet processes proposed by Teh et al. as a prior distribution of the model to flexibly discover more hidden states as additional narratives are observed [Teh *et al.*, 2006]. All narratives in each experiment are used to learn such a hierarchically-structured dynamical model.

We utilize the hierarchical prior in the following specification based on our problem scenario. Let G denote the global measure of an experiment (I or II), and it is distributed as

$DP(\gamma, G_0)$ with G_0 the base measure and γ the concentration parameter. Each π_k is conditionally independent given G. This hierarchical construction can be formulated as,

$$G \mid G_0 \sim DP(\gamma, G_0) \tag{1}$$

$$\begin{gathered} \pi_k \mid G \sim DP(\alpha, G) \\ k = 1, 2, \ldots, \infty \end{gathered} \tag{2}$$

In the i^{th} narrative, each transition probability distribution $\{\pi_{z_{t-1}, z_t = k}\}_{k=1,2,\ldots,\infty}$ of the hidden Markov model at the lower level governs the transitions toward hidden states ϕ_k's.

$$z_t^{(i)} \mid z_{t-1}^{(i)}, \pi_{z_{t-1}} \sim \pi_{z_{t-1}} \tag{3}$$

$$x_t^{(i)} \mid z_t^{(i)}, \phi_{z_t} \sim F(\phi_{z_t}) \tag{4}$$

4.4 Inference Algorithm

We use Markov chain Monte Carlo sampler to do the posterior inference over this model. In one iteration of the sampler, each latent variable is visited and assigned a value by drawing from the distribution of that variable conditional on the assignments to all other latent variables as well as the observation. In particular, based on the sampling algorithm proposed in [Van Gael *et al.*, 2008], we develop a sampling solution that uses multiple concept sequences with arbitrary lengths as observations. Specifically for each concept sequence $\{x_t^{(i)}\}_{t=1,2,\ldots,T_i}$, auxiliary variables $\{u_t^{(i)}\}$ are sampled with probability density,

$$\begin{gathered} p(u_t^{(i)} | z_{t-1}^{(i)}, z_t^{(i)}, \boldsymbol{\pi}) = \frac{\delta(0 < u_t^{(i)} < \pi_{z_{t-1}^{(i)}, z_t^{(i)}})}{\pi_{z_{t-1}^{(i)}, z_t^{(i)}}} \\ \delta(C) = \begin{cases} 1, \text{ if } C \text{ is true} \\ 0, \text{ otherwise} \end{cases} \end{gathered} \tag{5}$$

where each $u_t^{(i)}$ serves as a dynamic threshold at $t^{(i)}$ to partition the probability distribution $\{\pi_{z_{t-1}, k}\}_{k=1,2,\ldots,\infty}$ into a finite set of entries larger than $u_t^{(i)}$ and an infinite set smaller than $u_t^{(i)}$. Only the states k's within the finite set are considered when sampling $z_t^{(i)}$ that transits out of state $z_{t-1}^{(i)}$ during dynamic programming. This reduces the number of potential states to consider and hence makes the inference efficient.

$$\begin{aligned} & p(z_t^{(i)} | x_{1:t}^{(i)}, u_{1:t}^{(i)}, \boldsymbol{\pi}, \boldsymbol{\phi}) \\ \propto & p(z_t^{(i)}, u_t^{(i)}, x_t^{(i)} | x_{1:t-1}^{(i)}, u_{1:t-1}^{(i)}, \boldsymbol{\pi}, \boldsymbol{\phi}) \\ = & \sum_{z_{t-1}^{(i)}} p(x_t^{(i)} | z_t^{(i)}, \boldsymbol{\phi}) p(u_t^{(i)} | z_{t-1}^{(i)}, z_t^{(i)}, \boldsymbol{\pi}) p(z_t^{(i)} | z_{t-1}^{(i)}, \boldsymbol{\pi}) \\ & p(z_{t-1}^{(i)} | x_{1:t-1}^{(i)}, u_{1:t-1}^{(i)}, \boldsymbol{\pi}, \boldsymbol{\phi}), \text{and by applying Eq. 5:} \\ = & p(x_t^{(i)} | z_t^{(i)}, \boldsymbol{\phi}) \sum_{z_{t-1}^{(i)} : \pi_{z_{t-1}^{(i)}, z_t^{(i)}} > u_t^{(i)}} p(z_{t-1}^{(i)} | x_{1:t-1}^{(i)}, u_{1:t-1}^{(i)}, \boldsymbol{\pi}, \boldsymbol{\phi}) \end{aligned} \tag{6}$$

Beside resampling the auxiliary variables $\{u_t^{(i)}\}$ and the state sequences $\{z_t^{(i)}\}$ in each iteration, the algorithm also resamples the shared DP base measure G, the hyper-parameters α and γ, the emission probabilities $\boldsymbol{\phi}$, and the transition probabilities $\boldsymbol{\pi}$. Specifically G is sampled proportional to an additional set of auxiliary variables $\{m_{\cdot k}\}_{k=1,2,\ldots,K}$, where each $m_{\kappa k}$ is independent of others given $\boldsymbol{z}$, $\boldsymbol{G}$, and $\boldsymbol{\alpha}$.

$$\begin{aligned} & G = (G_1 \ldots G_K, \sum_{k'=K+1}^{\infty} G_{k'}) \sim \text{Dir}(m_{\cdot 1} \ldots m_{\cdot K}, \gamma) \\ & m_{\cdot k} = \sum_{\kappa=1}^{K} m_{\kappa k} \\ & p(m_{\kappa k} = m | \boldsymbol{z}, G, \alpha) \propto S(n_{\kappa k}, m)(\alpha G_k)^m \end{aligned} \tag{7}$$

where $S(\cdot, \cdot)$ denotes Stirling numbers of the first kind. Summing over the infinite many states that never occur in any hidden state sequences $\{z_t^{(i)}\}$, the conditional distribution $\pi_{k\cdot}$ given its Markov blanket $\boldsymbol{z}$, $\boldsymbol{G}$, and $\boldsymbol{\alpha}$ is

$$\begin{aligned} & \pi_{k\cdot} = (\pi_{k1} \ldots \pi_{kK}, \sum_{k'=K+1}^{\infty} \pi_{kk'}) \\ & \sim \text{Dir}(\sum_i n_{k1}^{(i)} + \alpha G_1 \ldots \sum_i n_{kK}^{(i)} + \alpha G_K, \alpha \sum_{k'=K+1}^{\infty} G_{k'}) \end{aligned} \tag{8}$$

where $n_{k\kappa}^{(i)}$ denotes the transition counts in the i-th state sequence from state k to κ. Each $\phi_{k\cdot}$ depends on the state sequences $\{z_t^{(i)}\}$, the observed concept sequences $\{x_t^{(i)}\}$, and the prior distribution $\boldsymbol{H}$, and the $\phi_{k\cdot}$'s are independent given $\boldsymbol{z}$, $\boldsymbol{x}$, and $\boldsymbol{H}$.

$$\phi_{k\cdot} \sim \text{Dir}(\sum_i l_{k1}^{(i)} + H_1 \ldots \sum_i l_{k|V|}^{(i)} + H_{|V|}) \tag{9}$$

where $l_{kv}^{(i)}$ denotes the emission counts in the i-th state sequence from state k to medical concept v. The whole vocabulary set is V. We further sample the hyper-parameters α and γ according to [Teh *et al.*, 2006].

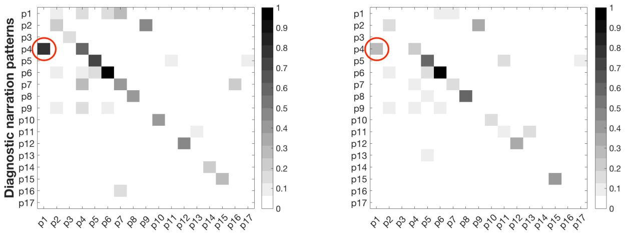
In each experiment we run the sampler 20 times with random initialization of the state sequences. Each state randomly chooses between 1 and the maximum length of all sequences. We use 2000 iterations as burn-in and empirically choose various hyperpriors for α and γ according to the convergence behaviors in previous runs. The hidden states inferred from the model are the diagnostic narration patterns mentioned in earlier sections, and *states* and *patterns* will be used interchangeably in the rest of this paper.

5 Results and Discussion

We summarize the state transitions within each narrative group and visualize the salient differences between groups. The involved states (patterns) appear frequently and they match the thought units. Classification results show that the diagnostic narration patterns contain key information in narratives to differentiate levels of diagnostic correctness.

5.1 The Discovered Verbal Narration Patterns

The state transition summaries for the two correctness levels in Figure 5 presents a salient difference—the state transition



(a) High-correctness group (b) Low-correctness group

Figure 5: Normalized state transitions in narrative groups regarding diagnostic correctness. One salient transition to discriminate both groups is from pattern 4 (the 4^{th} row) to 1 (the 1^{st} column).

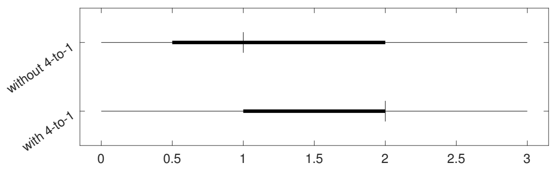
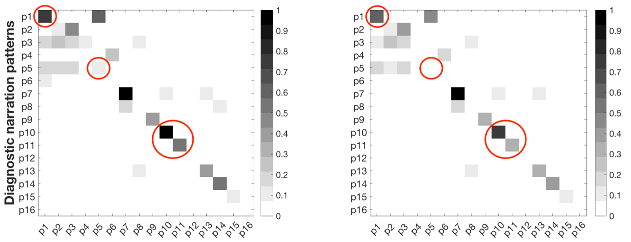


Figure 6: The correctness score distributions between the narratives with state transition (4 → 1) and those without.



(a) High-confidence group (b) Low-confidence group

Figure 7: Normalized state transitions in narrative groups regarding diagnostic confidence. Group (a) possesses slightly more self-transitions of 1, 5, 10 and 11 than (b).

from pattern 4 to 1. These two patterns are interpretable—Pattern 1 can be interpreted as *primary lesion type* (PRI), and pattern 4 includes informative findings regarding color, size, shape and texture of the lesion to assist determining the primary lesion type. Given the meanings of these patterns, we find that the high-correctness narratives possess more transitions from describing supportive findings to mentioning primary lesion type than the low-correctness narratives. We also consider all the narratives in Experiment I, and separately visualize the ones with and without state transition $4 \rightarrow 1$ in Figure 6. We notice two different distributions of correctness scores, and we find that the narratives with this key state transition generally shift towards the higher correctness end. This implies the importance of locating key clues before determining a primary lesion type in order to make a correct diagnosis. Similar patterns are discovered in parallel from the diagnostic confidence study in Experiment II. Since these patterns appear in both experiments and match the thought units, we recognize them as *Signature Patterns*. Diagnostic confidence study also presents different state transitions between low and high confidence levels, which involve interpretable patterns (see Figure 7)—Patterns 1 and 5 presents primary lesion type and informative findings. Pattern 10 can be interpreted as *confidence marker*, and pattern 11 as *diagnosis* or *differential diagnoses*.

5.2 Narrative Correctness Classification

We classify the narratives at low and high correctness levels based on various feature combinations with two classifiers, and Table 2 summarizes the classification performances. We use cross-validation to tune the trade-off parameter of the lasso-regularized logistic regression. Cross-validation is also used to determine the optimal number of hidden states for the canonical HMM. We find that the infinite HMM works better than the canonical HMM to capture the important temporal patterns for classification. Concatenating all features boosts the classification performance, as the LDA and M-seq iHMM capture high-level information complementary to the fine details in Bag-of-Word (TF-IDF). Both classifiers suggest high importance of the patterns learned from M-seq iHMM, and Table 3 lists highest-ranked features and their interpretations.

In Figure 8, we analyze both high-correctness narratives (B and D) and low ones (A and C). The narratives (A) and (B) are successfully classified. (C) is misclassified as high-correctness, because it mentions the correct primary lesion type but fails to give a correct diagnosis nor differential diagnoses. (D) is misclassified as low-correctness, because it only makes a correct diagnosis without mentioning the primary lesion type. These examples show that the signature patterns dominate the classifier, and hereby fail to capture the correctness related to diagnoses. This is because Experiment I covers 46 diagnoses by 48 images, which makes each disease name appear infrequently. We omit the diagnostic confidence classification due to space limits.

6 Conclusions

This paper explores diagnostic decision-making by modeling physicians' utterances of medical concepts during image-based diagnoses. We develop automated approaches to discover diagnostic narration patterns from expert data. Our model discovers patterns that exist in both datasets and match the expert-defined diagnostic thought units. These patterns are also important features for diagnostic correctness classification. The approaches proposed in this paper can facilitate education in the medical fields, research in cognition and decision-making, and medical image classification based on physicians' thoughts.

Since the concepts in the same narrative are essentially correlated, we plan to relax the strong Markovian assumption in the model by considering semantic relatedness of medical concepts [Liu *et al.*, 2012; Bollegala *et al.*, 2015] and developing an additional variant in the future. We will further explore medical image difficulty levels [Guo *et al.*, 2014] and physicians' expertise levels [Wu *et al.*, 2015] as they are relevant factors in diagnostic decision making and error prevention.

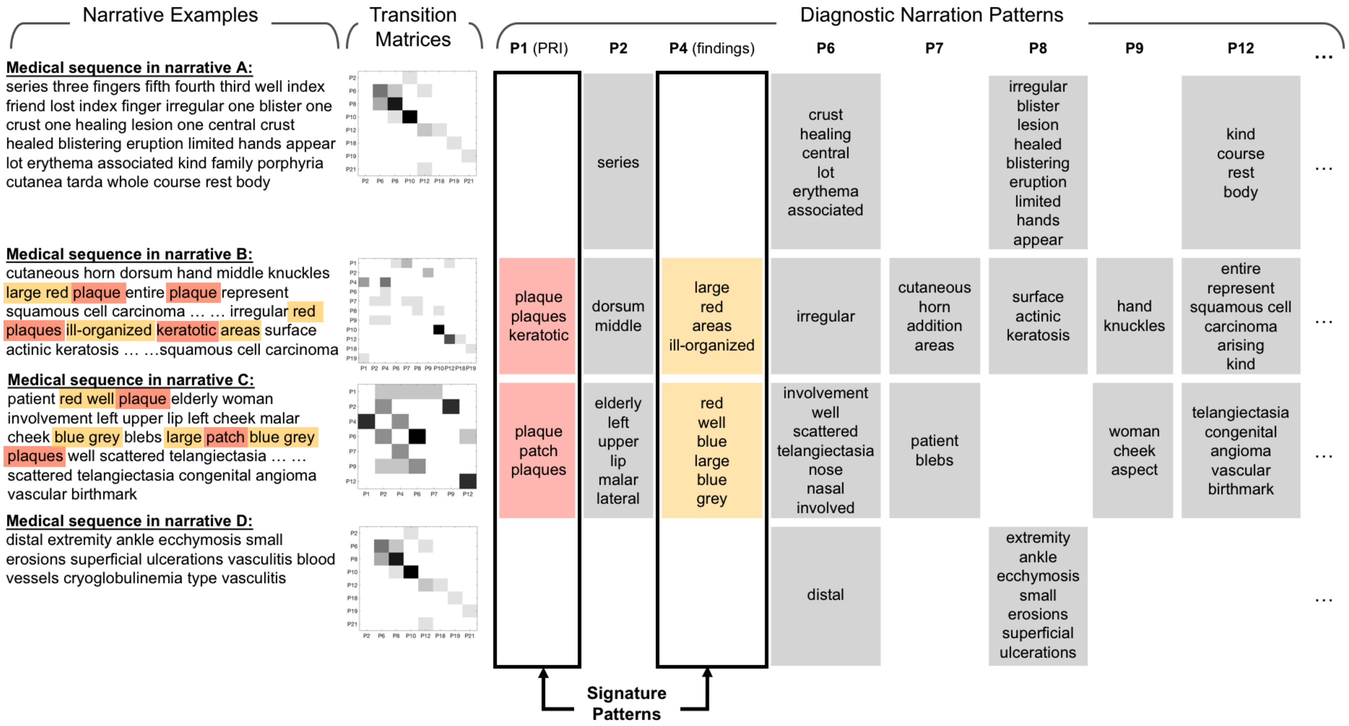

Narrative Examples
Transition Matrices
Diagnostic Narration Patterns
Medical sequence in narrative A:
series three fingers fifth fourth third well index friend lost index finger irregular one blister one crust one healing lesion one central crust healed blistering eruption limited hands appear lot erythema associated kind family porphyria cutanea tarda whole course rest body
Medical sequence in narrative B:
cutaneous horn dorsum hand middle knuckles large red plaque entire plaque represent squamous cell carcinoma … … irregular red plaques ill-organized keratotic areas surface actinic keratosis … …squamous cell carcinoma
Medical sequence in narrative C:
patient red well plaque elderly woman involvement left upper lip left cheek malar cheek blue grey blebs large patch blue grey plaques well scattered telangiectasia … … scattered telangiectasia congenital angioma vascular birthmark
Medical sequence in narrative D:
distal extremity ankle ecchymosis small erosions superficial ulcerations vasculitis blood vessels cryoglobulinemia type vasculitis
P1 (PRI)
P2
P4 (findings)
P6
P7
P8
P9
P12
…
series
crust healing central lot erythema associated
irregular blister lesion healed blistering eruption limited hands appear
kind course rest body
plaque plaques keratotic
dorsum middle
large red areas ill-organized
irregular
cutaneous horn addition areas
surface actinic keratosis
hand knuckles
entire represent squamous cell carcinoma arising kind
plaque patch plaques
elderly left upper lip malar lateral
red well blue large blue grey
involvement well scattered telangiectasia nose nasal involved
patient blebs
woman cheek aspect
telangiectasia congenital angioma vascular birthmark
distal
extremity ankle ecchymosis small erosions superficial ulcerations
Signature Patterns

Figure 8: Example narratives in the diagnostic correctness study. *Left*: the remaining medical concept sequences in four narrative examples. *Middle*: the corresponding transition probability matrices out of the overall 17 patterns discovered from all narratives in Experiment I. *Right*: the shared narration pattern matrix. The two *Signature Patterns* are highlighted.

Classifier / Feature	Regularized Logistic Regression		Random Forest	
	Accuracy (%)	AUC of ROC	Accuracy (%)	AUC of ROC
TF-IDF	61.8	0.63	44.9	0.65
LDA	64.0	0.63	62.9	0.67
HMM	59.6	0.62	56.2	0.58
M-seq iHMM	64.0	0.68	65.2	0.64
TF-IDF + M-seq iHMM	**67.4**	0.69	**75.3**	**0.78**
TF-IDF + LDA + M-seq iHMM	**67.4**	**0.71**	**75.3**	**0.78**

Table 2: Narrative correctness classification performances. The positive class for ROC is high-correctness.

Rank	Feature (Feature interpretations in detail)
1	Pattern 4 in M-seq iHMM (erythemas, pinch purpura, annulare, ...)
2	Topic 45 in LDA (hand, dorsal, hyperkeratotic, ...)
3	Term 32 in TF-IDF (papules)
. . .	. . .

Table 3: Ranked features by random forest classifier.

Acknowledgments

This work was partially supported by NIH Grant 1R21 LM010039-01A1 and NSF grant IIS-0941452. We would like to thank the participating physicians, the reviewers, and Logical Images, Inc. for images. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the official views of the NIH or the NSF.

References

[Aronson, 2001] Alan R Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: The Metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[Berner and Graber, 2008] Eta S Berner and Mark L Graber. Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, 121(5):S2–S23, 2008.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[Boersma and Weenink, 2009] Paul Boersma and David Weenink. Praat: Doing phonetics by computer (version 5.1. 05) [computer program]. `http://www.praat.org/` (Accessed: 10 April 2014), 2009.

[Bollegala *et al.*, 2015] Danushka Bollegala, Takanori Maehara, and Ken-ichi Kawarabayashi. Embedding semantic relations into word representations. In *Proceedings of the* 24^{th} *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1222–1228. AAAI Press, 2015.

[Bowen, 2006] Judith L Bowen. Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355(21):2217–2225, 2006.

[Croskerry, 2009] Pat Croskerry. A universal model of diagnostic reasoning. *Academic Medicine*, 84(8):1022–1028, 2009.

[Deerwester *et al.*, 1990] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391, 1990.

[Fung and Bodenreider, 2012] Kin Wah Fung and Olivier Bodenreider. Knowledge representation and ontologies. In *Clinical Research Informatics*, pages 255–275. Springer, 2012.

[Galanter and Patel, 2005] Cathryn A Galanter and Vimla L Patel. Medical decision making: A selective review for child psychiatrists and psychologists. *Journal of Child Psychology and Psychiatry*, 46(7):675–689, 2005.

[Guo *et al.*, 2014] Xuan Guo, Qi Yu, Cecilia Ovesdotter Alm, Cara Calvelli, Jeff B Pelz, Pengcheng Shi, and Anne R Haake. From spoken narratives to domain knowledge: Mining linguistic data for medical image understanding. *Artificial Intelligence in Medicine*, 62(2):79–90, 2014.

[Liu *et al.*, 2012] Ying Liu, Bridget T McInnes, Ted Pedersen, Genevieve Melton-Meaux, and Serguei Pakhomov. Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet. In *Proceedings of the* 2^{nd} *ACM SIGHIT International Health Informatics Symposium*, pages 363–372. ACM, 2012.

[Lyons and Ousley, 2014] Faye Lyons and Lisa Ellen Ousley. *Dermatology for the advanced practice nurse*. Springer Publishing Company, 2014.

[McCoy *et al.*, 2012] Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Rui Li, Jeff B Pelz, Pengcheng Shi, and Anne Haake. Annotation schemes to encode domain knowledge in medical narratives. In *Proceedings of the* 6^{th} *Linguistic Annotation Workshop*, pages 95–103, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[Rabiner, 1989] Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[Teh *et al.*, 2006] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 2006.

[Van Gael *et al.*, 2008] Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. Beam sampling for the infinite hidden Markov model. In *Proceedings of the* 25^{th} *International Conference on Machine Learning*, pages 1088–1095. ACM, 2008.

[Wu *et al.*, 2015] Run-ze Wu, Qi Liu, Yuping Liu, Enhong Chen, Yu Su, Zhigang Chen, and Guoping Hu. Cognitive modeling for predicting examinee performance. In *Proceedings of the* 24^{th} *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1017–1024. AAAI Press, 2015.