LATENT SEMANTIC MODELING FOR SLOT FILLING IN CONVERSATIONAL UNDERSTANDING

Gokhan Tur Asli Celikyilmaz Dilek Hakkani-Tür

Microsoft Silicon Valley

ABSTRACT

In this paper, we propose a new framework for semantic template filling in a conversational understanding (CU) system. Our method decomposes the task into two steps: latent n-gram clustering using a semi-supervised latent Dirichlet allocation (LDA) and sequence tagging for learning semantic structures in a CU system. Latent semantic modeling has been investigated to improve many natural language processing tasks such as syntactic parsing or topic tracking. However, due to several complexity problems caused by issues involving utterance length or dialog corpus size, it has not been analyzed directly for semantic parsing tasks. In this paper, we propose extending the LDA by introducing prior knowledge we obtain from semantic knowledge bases. Then, the topic posteriors obtained from the new LDA model are used as additional constraints to a sequence learning model for the semantic template filling task. The experimental results show significant performance gains on semantic slot filling models when features from latent semantic models are used in a conditional random field (CRF).

Index Terms— spoken language understanding, slot filling, latent semantic modeling, graphical models

1. INTRODUCTION

Spoken language understanding (SLU) aims to extract the *meaning* of speech utterances. More specifically, targeted SLU models in human/machine spoken dialog systems aim to automatically identify several components: (*i*) the domain and intent of the user utterance as expressed in natural language, (*ii*) the slots, associated arguments, attributed to phrases in the utterance [1]. The aim is to pass these semantic components to the dialog engine in order to achieve a certain task, e.g., query a database for the list of movies that users request. An example output of a parsed utterance from a movies domain is shown in Table 1. A common approach to semantic parsing in SLU is using a classification method for filling frame slots given an application domain. These approaches include generative models such as hidden Markov models [2], discriminative methods [3, 4, 5], or probabilistic context free grammars [6, 7], to name a few.

In this paper, our goal is two-fold: (*i*) discovering correlated terms or phrases of a given domain from in-domain unlabeled utterances as well as large resources of unstructured text collected from the web (e.g., reviews or blogs on movies, restaurants, etc.) and semantic knowledge bases; (*ii*) improving the slot filling task by generalizing from a smaller corpus, which is labeled with domain specific slot types. One of the challenges of our task is collecting labeled corpora for each domain, which is tedious and noise prone. We claim that generalizing terms with correlated meanings, and later injecting them as additional constraints for slot filling, may enrich the feature set. For example, the word "funny" is typically used to describe a movie, and is annotated in the labeled data as a "movie description" tag. In this paper, we use unsupervised clustering on large unlabeled online documents to discover semantically similar words, e.g., given that "*funny*" exists in training data we discover semantically similar words/phrases such as "*hilarious*", "*made me laugh for hours*", etc. which are also used to describe movies. Once we extend the slot value lists, we use them as additional information for the slot filling task. Similar generalizations may also be made for terms forming the lexical context of specific slots (e.g., "directed by *director-name*").

Utterance	*show me recent action movies by spielberg*
Domain:	Movie
Intent:	Find_Movie
Genre:	*action*
Date:	*recent*
Director:	*spielberg*

Table 1. An example utterance with semantic annotations.

This paper proposes a generic and theoretically sound mechanism for understanding natural language utterances that goes beyond local lexical features but rather enables longer dependencies using utterance level features for semantic tagging. We use Latent Dirichlet Allocation (LDA) [8] to capture the correlated terms in given documents. Recent work has used topic models for natural language proessing (NLP) tasks including statistical analysis of document collections. For instance, in a recent work [9, 10] used unsupervised latent variable models to cluster utterances into semantic clusters using Bayesian inference. A classical LDA assumes a range of possible distributions, constrained by being drawn from Dirichlet distributions. This enables a latent topic model to be learned entirely unsupervised, and allows the model to be maximally relevant to the data being segmented.

Although shown to improve many NLP tasks, topic models can help improve other tasks better when some supervision is provided to the algorithm, e.g., in semantic slot filling, prior information might be in the form of correspondence between a latent topic and one or more of the semantic slot types. In fact, the semantic modeling research community has recently investigated the use of prior information in latent topic models to preserve one-to-one correspondence between the latent topics and labeled semantic components. For instance, [11] presented the Labeled LDA model, which captures the latent topics that correspond to the user tags and applied them to text classification problems. Similarly, [12] introduced a new topic model, the Distance Dependent Semi-Latent Topic Model (dd-SLDA), to capture latent topics from related utterances in CU systems and applied their model to the dialog act (intent) detection problem. They defined the dialog acts as hidden aspects of utterances and used intent labeled utterances to assign each semantic cluster to one of the set of predefined intent clusters. On the speech processing side, latent semantic models were first employed by Bellegarda [13], for training semantic language models. In their approach, the (dis-

crete) words and documents are mapped onto a (continuous) semantic vector space (other clustering techniques have also been applied for language modeling). Another recent work on speech data has explored topic detection of spoken documents using LDA-style graphical models [14]. Nevertheless, there has not been much focus on latent semantic modeling approach to semantic slot detection.

In this paper, we tackle the problem of semantic component extraction from utterances, namely semantic slot mapping. Thus, we take each semantic tag or slot type as a latent aspect of utterances. We use topic clustering on unlabeled text (reviews, blogs, utterances, etc) to discover latent topic clusters of semantic slot types. We extend LDA using gazetteers extracted from knowledge bases as prior information at training time. Specifically, when generating words, we use the slot-type information and generate multiple topics for each slot type to preserve slot-topic relations. The slot type can be provided using indirect supervision from pre-compiled gazetteers (such as lists of genre types). We start with a completely unsupervised LDA model and incrementally add this prior information during learning the parameters. We show on a test set that using seed labeled data to capture slot posteriors for utterances through a latent variable model significantly improves the slot filling performance.

The next section describes the generic problem of slot filling for CU, and very briefly presents the state-of-the-art discriminative classification approach using conditional random fields (CRFs). In Section 3, we provide a high level overview of the latent semantic modeling, and we describe how it can be used for improving the slot filling task via CRFs. Experimental results are presented using a representative CU system in Section 4.

2. SEMANTIC PARSING

Following the state-of-the-art approaches for slot filling [4, 5, among others], we use discriminative statistical models, namely conditional random fields, (CRFs) [15], for modeling. More specifically and formally, slot filling is framed as a sequence classification problem to obtain the most probable slot sequence:

$$\hat{Y} = \underset{Y}{\operatorname{argmax}}\, p(Y|X)$$

where $X = x_1, \ldots, x_T$ is the word sequence and $Y = y_1, \ldots, y_T$, $y_i \in C$ is the sequence of associated class labels, C.

CRFs are shown to outperform other classification methods for sequence classification [16], since the training can be done discriminatively over a sequence with sentence level optimization. The baseline model relies on a word n-gram based linear chain CRF, imposing the first order Markov constraint on the model topology. Similar to maximum entropy models, in this model, the conditional probability, $p(Y|X)$ is defined as [15]:

$$p(Y|X) = \frac{1}{Z(X)} exp\left(\sum_k \lambda_k f_k(y_{t-1}, y_t, x_t)\right)$$

with the difference that both X and Y are sequences instead of individual local decision points given a set of features f_k (such as n-gram lexical features, state transition features, or others) with associated weights λ_k. $Z(X)$ is the normalization term. After the transition and emission probabilities are optimized, the most probable state sequence, $\hat{Y}$, can be determined using the well-known Viterbi algorithm.

3. LATENT SEMANTIC MODELING FOR SLOT FILLING

In literature there have been different approaches to Latent Semantic Models, which are general techniques in the NLP world. They mainly analyze the relationship between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. In Latent Semantic Analysis (LSA), or Latent Semantic Indexing (LSI) [17], it is assumed that the words which are close in meaning will occur in similar pieces of text. A matrix containing word counts per paragraph is constructed from a large piece of text and a mathematical technique called singular value decomposition (SVD) is used to reduce the number of columns while preserving the similarity structure among rows. LSA cannot capture polysemy (i.e., multiple meanings of a word) and hence each occurrence of a word is treated as having the same meaning due to the word being represented as a single point in space. To overcome this problem, probabilistic latent semantic analysis (PLSA) [18], also known as probabilistic latent semantic indexing (PLSI), is introduced. Using PLSA one can derive a low dimensional representation of the observed variables in terms of their affinity to certain hidden variables, just as in latent semantic analysis. PLSA evolved from latent semantic analysis, adding a sounder probabilistic model. PLSA, however is not a complete graphical model for new documents, since a new model should be trained as a new document is introduced. Thus LDA models have been introduced to overcome these problems of PLSA. Next we briefly explain the LDA learning algorithm and our extension, namely the semi-supervised LDA. Later we present new scores - which utilize the posteriors obtained from these trained topics models - as constraints for predicting slot posteriors in given utterances.

3.1. Latent Dirichlet Allocation (LDA)

LDA is an admixture model where the documents are modeled as distributions over sets of hidden topics and each hidden topic is also considered to be a distribution over words in the corpus. The model assumes that there are K underlying topics, according to which documents are generated.

A document is generated by sampling a mixture of the semantic classes (topics) and then sampling word n-grams conditioned on a particular semantic class. Each document is assumed to be drawn from a mixture of K shared topics, with topic z receiving a weight $\theta_z^{(u)}$ in document u. Each topic is a distribution over a shared vocabulary of W words, with each word w having probability $\phi_w^{(z)}$ in topic z. Dirichlet priors are used to regularize θ and ϕ. The generative process of the `LDA` model (Fig. 1 *left*) can be formalized as:

1. Choose $\theta^{(u)} \sim Dir(\alpha)$, *u=1,...,*$|U|$, and choose $\phi^{(z)} \sim Dir(\beta)$, $z = 1, \ldots, K$.
2. For each word $w_{u,n}$ in each document u:
   (a) Choose a topic $z_n \sim Mult(\theta^{(u_n)})$
   (b) Choose a word $w_n \sim \phi^{(z_n)}$

The α and β are fixed hyper-parameters and we need to estimate parameters θ for each document and ϕ for each topic. From the expectation of the Dirichlet distributions, the probability of a document $\mathbf{u}$=$w_1, \ldots, w_{N_u}$ is given by:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N_u} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta)\right) d\theta \quad (1)$$

Gibbs sampling is one of the practical solutions for Bayesian inference and collapsed Gibbs sampling is a variant where two random variables, the θ, ϕ, are analytically integrated out.

In LDA, the posterior probability of the topic label z_i for word i, conditioned on the rest of the words 1 to $n-1$ and their topic labels,

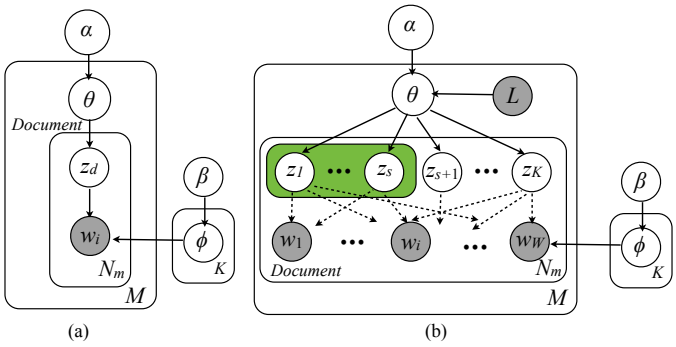

Fig. 1. Graphical model depiction of: (a) LDA; (b) Semi-Supervised LDA (SSLDA). Blank circles indicate latent variables, whereas dark-gray filled circles indicate known variables. L is the prior knowledge injected as binary lattice (L as shown as known variable, d is the number of latent topics corresponding to known slot clusters.

is formulated as:

$$P(z_i|z_{n\setminus i}, w_n) \propto \frac{n^{(w_i)}_{z_i,n\setminus i} + \beta}{n^{(\cdot)}_{z_i,n\setminus i} + W\beta} \cdot \frac{n^{(u_i)}_{z_i,n\setminus i} + \alpha}{n^{(u_i)}_{\cdot,n\setminus i} + K\alpha} \quad (2)$$

where $n^{(w_i)}_{z_i,n\setminus i}$ is the number of words assigned to topic i that are the same as w, $n^{(\cdot)}_{z_i,n\setminus i}$ is the total number of words assigned to topic i, $n^{(u_i)}_{z_i,n\setminus i}$ is the number of words from document u assigned to topic i, and $n^{(u_i)}_{\cdot,n\setminus i}$ is the total number of words in document u. $.\setminus i$ indicates counts that do not include the item i.

3.2. Context-Aware Document Generation for Topic Models

Mixture modeling of documents into topical semantic clusters is proven to be effective for SLU tasks, where the goal is finding the global aspects such as domain, topic, or intent of a given utterance (e.g., [9, 12]). In this study, however, we have a different challenge: Rather than classifying utterances, we classify words in sequence and assign a slot tag. Our focus is mainly on semantic clustering of words along with their context information. Hence, rather than using utterances as documents, for each word, we compile documents based on their context and inject "*direct*" and "*indirect*" supervision. For example, a typical utterance sequence can be composed of word n-grams like "*schedule*", "*3 pm*", "*cafe plaza*", etc., each of which may correspond to different semantic topics corresponding to a specific slot type, e.g., type, time, location.

In order to achieve this goal, we create *pseudo-documents* for each word, with their lexical contexts. With this method of transforming the word-document matrix into context-word matrix, the words (documents for LDA) with similar contexts (words for LDA) would be clustered together. More formally, following the above notation, the lexicon of LDA is now the list of possible contexts for each word w_i. As context, we employ (one or two) previous (L) and/or next (R) words:

$L : w_{i-1}, R : w_{i+1}, LR : w_{i-1}\_w_{i+1},$

$LL : w_{i-2}\_w_{i-1}, RR : w_{i+1}\_w_{i+2}$

Each word is assumed to be drawn from a mixture of K shared topics, with topic z receiving a weight $\theta^{(u)}_z$ in word u. Each topic is a distribution over a shared vocabulary of W contexts, with each context w having probability $\phi^{(z)}_w$ in topic z.

3.3. Semi-Supervised Latent Dirichlet Allocation (SSLDA)

We now turn our attention to our proposed approach where we inject prior knowledge into an LDA model as labeled latent topics. We use the context aware documents of a given word to build the probabilistic model. Specifically, the document structures defined below correspond to the pseudo-documents for each context-word as explained in the previous section.

In a semantic slot tagging task, we would like to attribute each n-gram (in a given document) to a possible semantic slot type. We also would like to build a more focused model, where there is a one-to-many map between the semantic slot classes and latent topics. To achieve this, we use an informative prior during Gibbs sampling, which pulls word-slot relations from lexicon dictionaries (namely gazetteers). Specifically, at training time, we provide a list of gazetteers, which we know *a priori* correspond to one or more slot types in our corpus. For example, the movie-genre dictionary items can fill slot values of *movie-genre* in the training data. For those documents of which we know the semantic slot type label of the context-word (based on a search in provided dictionaries), we sample the words from the topics designated for that semantic class, namely topics corresponding to slot types. Similarly, for the unlabeled documents whose semantic slot tags are not known, we sample topics of each word n-gram as follows: if an unlabeled word exists in one or more lexicon dictionaries, we introduce the prior belief that this word should be emitted by the slot types that those two lexicon dictionaries correspond to. Similarly, if the word does not exist in any of the lexicons, we let the algorithm decide which topic that word should belong to.

Thus, at training time, we construct a lattice of lexicon-topic-words to be used as prior information. During model training and inference, we use this lattice as restrictive information when generating each word in each document. We reserve s number of latent topics $z_1, \ldots, z_s$ to sustain a correspondence between the latent topics and the semantic labels (slot types) as shown in the graph representation of SSLDA in (Fig. 1 *right*). The rest of the topics may or may not correspond to any slot type in our corpus.

A set of documents D is a vector of N_d ngrams, $\mathbf{w}_d = \{w_{nd}\}_{n=1}^{N_d}$, where each $w_{nd} \in \{1, \ldots, V\}$ is chosen from a vocabulary of size V, and a vector of s slots, chosen from a set of semantic classes of size S.

Step-1 Designate the first s topics to the known slot types of the training dataset. Generate a binary lattice $\mathcal{L}_{w\times s}$ of word versus slot types using the lexicon dictionaries.

Step-2: Build a semi-supervised LDA (SSLDA) model on sets of documents D. This process is similar to the LDA except that when sampling words for a document, whose slot is known a priori, we sample from the first s topics that are designated for that semantic class (slot). The generative process of the graphical model can be formalized as:

1. Choose $\theta^{(d)} \sim Dir(\alpha)$, *d=1,…,*$|\mathcal{D}|$, and choose $\phi^{(z)} \sim Dir(\beta)$, $z = 1, \ldots, K$.
2. For each word n-grams $w_{d,n}$ in each utterance $\mathcal{D}$:
   (a) Find possible slot $s_{w_{d,n}}$ for the $w_{d,n}$ based on the $\mathcal{L}_{w_{d,n}\times s}$ and later sample a topic $z_{s_n} \sim Mult(\theta^{(d_n)})$ only from those topics containing that word. If the word does not exist on any of the possible topic lexicons, sample a $z_k \sim Mult(\theta^{(d_n)})$ from any topic.
   (b) Choose a word n-gram $w_n \sim \phi^{(z_{s_n}, s_{w,d_n})}$

A topic is sampled to generate each n-gram using:

$$p(z = k|w_n, s, \mathbf{z}_{-i}) = P(z_i|z_{\mathbf{s}\setminus\mathbf{i}}, w_n) * I[w_{d,n} \in \tilde{s}_{w_{d,n}}] \quad (3)$$

	No. Utt.	Avg. No. Words	Avg. No. Slots
Seed Training	3,454	5.73	1.88
All Training	22,677	5.06	-
Test	5,880	4.78	1.49

Table 2. Data sets used in the experiments.

Model	All	Unnamed Slots	Named Slots
Baseline	73.53%	77.80%	67.00%
k-means	74.15%	79.77%	67.81%
LDA	**75.79%**	**79.52%**	**71.26%**

Table 3. Experimental results for exploiting unsupervised latent semantic information.

The indicator, $I[.]$, is used to eliminate those slots that the word n-gram $w_{d,n}$ has not been identified in the lattice $\mathcal{L}_{w_{d,n} \times d}$, hence the designated topics are not sampled from them. Instead of using random topic sampling, e.g., uninformative prior of unsupervised LDA, we use an informative prior that preferentially assigns a given word to topics that this word has been associated with before. For instance, if the $w_{d,n}$ has been used as part of actor name and director name slots, it is very likely that one of these slots will be chosen as the topic, z_s.

3.4. Using Latent Variables for Slot Filling

In this paper, we focus on a semantic slot detection problem of a conversational understanding engine, for a representative domain: movies. Hence, our labeled utterances are collections from users' interactions with computers or mobile devices where they would like to get more information about movies, find out about showtimes, places and more information about the cast. Hence we use two datasets to build the graphical models on: To bootstrap, we used a small set of annotated examples, and we used a larger unannotated dataset for semi-supervised classification.

First annotated and unannotated sets are concatenated to train LDA models. Then these context sensitive clusters are applied for CRF modeling in a straightforward manner. The IOB schema is adopted, following the literature, where each of the words are tagged with their position in the slot: beginning (B), in (I) or other (O). For each word in the small annotated set, the most probable cluster in that context is provided as an additional feature, $\text{argmax}_z\, \theta_z^{(u)}$ for each word u. Exploiting its score or cluster score distributions are left as future research.

4. EXPERIMENTS AND RESULTS

Experiments are performed using an SLU system, with real users. The users present queries about various movies, such as "*who is the director of avatar*", "*show me some action movies with academy awards*", or "*when is the next harry potter gonna be released*". The semantic space consists of 26 slot types, such as named ones (movie or actor names) or unnamed ones (genre or language). Table 2 shows the properties of the data sets. Only a small portion of the training data is manually annotated with semantic slots.

The knowledge base (similar to Freebase) for the movies domain is used to mine weighted gazetteers for 5 slot types: genre, language, nationality, MPAA-rating, and release-date. These are weighted with respect to their prior probability in the knowledge base.

For evaluation, the slot F-measure is used, following the literature [5] using the CoNLL evaluation script[1]. The baseline performance is obtained using only word n-grams with a linear chain CRF using the CRF++ toolkit[2] using default parameters with word level IOB format. The number of clusters (K) is always set to 100.

foreign,british,spanish,indian
comedies,ones,dramas,documentaries
want,had,understand,know
weeds,scrubs,dexter,rango
four,eight,seven,six,five
they,i,we,he,she
seventy,ninety,eighty,sixty,season
hanks,cruise,newman,ford,stiller
set,made,filmed,released,available
synopsis,bios,summary,details
robert,brad,steven,sean,patrick

Table 4. Most probable words in most frequent clusters with LDA.

Model	All	Unnamed Slots	Named Slots
Baseline	74.53%	80.79%	66.97%
k-means	73.29%	81.30%	66.66%
SSLDA	**76.49%**	**81.33%**	**72.90%**

Table 5. Experimental results for exploiting semi-supervised latent semantic information.

In order to compare the effectiveness of LDA with other simpler methods, we have also implemented a k-means clustering algorithm, an EM based approach, where each word is iteratively assigned to a more similar cluster as described in [19].

Table 3 presents the results using only lexical features without supervision during graphical modeling. This experiment shows the added value of unsupervised semantic clustering for the task of slot filling. The use of latent semantic information significantly[3] improves the slot filling performance from 73.53% to 75.79%. When we look at slot-level performances, we see that k-means and LDA both improve unnamed slots but LDA is also effective for named slots, a big differentiator to the k-means clustering method.

The second batch of experiments employs light or indirect supervision during graphical modeling only for the 5 unnamed slot types listed above. Table 4 presents most probable words in most frequent clusters, very informative for slot filling.

Table 5 presents the results with prior knowledge as obtained from gazetteers. Note that these 5 gazetteers for 5 unnamed slot types have also been used during CRF training as additional features to perform more fair experiments. This resulted in about 3% F-measure improvement for these slots. While k-means clustering results in slight improvement on top of this, the semi-supervised LDA approach performs the best, reaching an overall F-measure of 76.49%, as the improvements are also propagated to named slot types similar to the experiments with unsupervised LDA.

5. CONCLUSIONS

We have presented a generic latent semantic slot filling modeling approach. While latent semantic models have been used in many NLP tasks, to the best of our knowledge, this is a pioneering study for slot filling, a key task in human/machine conversational systems.

The context sensitive clustering approach naturally suggests semi-Markov CRF modeling, instead of linear CRF. We plan to experiment using that schema in our future research, exploiting all cluster score distributions.

[1] http://www.cnts.ua.ac.be/conll2000/chunking/output.html

[2] http://crfpp.sourceforge.net

[3] According to the McNemar significance test [20], $p < 0.001$

6. REFERENCES

[1] G. Tur and L. Deng, *Intent Determination and Spoken Utterance Classification, Chpater 3 in Book: Spoken Language Understanding*, John Wiley and Sons, New York, NY, 2011.

[2] R. Pieraccini, E. Tzoukermann, Z. Gorelov, J.-L. Gauvain, E. Levin, C.-H. Lee, and J. G. Wilpon, "A speech understanding system based on statistical representation of semantics," in *Proceedings of the ICASSP*, San Francisco, CA, March 1992.

[3] R. Kuhn and R. De Mori, "The application of semantic classification trees to natural language understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 449–460, 1995.

[4] Y.-Y. Wang and A. Acero, "Discriminative models for spoken language understanding," in *Proceedings of the ICSLP*, Pittsburgh, PA, September 2006.

[5] C. Raymond and G. Riccardi, "Generative and discriminative algorithms for spoken language understanding," in *Proceedings of the Interspeech*, Antwerp, Belgium, 2007.

[6] S. Seneff, "TINA: A natural language system for spoken language applications," *Computational Linguistics*, vol. 18, no. 1, pp. 61–86, 1992.

[7] W. Ward and S.Issar, "Recent improvements in the CMU spoken language understanding system," in *Proceedings of the ARPA HLT Workshop*, March 1994, pp. 213–216.

[8] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, 2003.

[9] A. Celikyilmaz, D. Hakkani-Tur, and G. Tur, "Multi-domain spoken language understanding with approximate inference," in *Proceedings of the Interspeech*, 2011.

[10] M. Dowman, T.L. Griffiths V. Savova, K.P. Krding, J.B. Tenenbaum, and M. Purver, "A probabilistic model of meetings that combines words and discourse features," in *IEEE Transactions on Audio, Speech and Language Processing*, 2008, vol. 16, pp. 1238–1248.

[11] R. Nallapati D. Ramage, D. Hall and C. Manning, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the NAACL*, 2009.

[12] A. Celikyilmaz, D. Hakkani-Tur, Gokhan Tur, A. Fidler, and D. Hillard, "Exploiting distance based similarity in topic models for user intent detection," in *Proceedings of the ASRU 2011*, Waikoloa, HI, 2011.

[13] J. R. Bellegarda, "A latent semantic analysis framework for large-span language modeling," in *Proceedings of the EUROSPEECH*, Rhodes, Greece, September 1997.

[14] T. J. Hazen, "Direct and latent modeling techniques for computing spoken document similarity," in *Proceedings of the IEEE SLT Workshop*, Berkeley, CA, 2010.

[15] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the ICML*, Williamstown, MA, 2001.

[16] G. Tur and R. De Mori, Eds., *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, John Wiley and Sons, New York, NY, 2011.

[17] S.T. Dumais, "Latent semantic analysis," in *Annual Review of Information Science and Technology*, 2005, vol. 38.

[18] T. Hoffman, "Learning the similarity of documents : an information-geometric approach to document retrieval and categorization," in *Advances in Neural Information Processing Systems*, 2000, vol. 112, pp. 914–920.

[19] J.P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt, "A hidden Markov model approach to text segmentation and event tracking," in *In Proceedings of the IEEE ICASSP*, Seattle, WA, May 1998.

[20] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proceedings of the ICASSP*, Glasgow, Scotland, 1989.