Approximate Inference for Domain Detection in Spoken Language Understanding

Asli Celikyilmaz[1], *Dilek Hakkani-Tür*[1],[2], *Gokhan Tür*[1],[2]

[1] Microsoft Speech Labs | [2]Microsoft Research
Mountain View, CA, 94041

asli|dilek|gokhan.tur@ieee.org

Abstract

This paper presents a semi-latent topic model for semantic domain detection in spoken language understanding systems. We use labeled utterance information to capture latent topics, which directly correspond to semantic domains. Additionally, we introduce an 'informative prior' for Bayesian inference that can simultaneously segment utterances of known domains into classes and divide them from out-of-domain utterances. We show that our model generalizes well on the task of classifying spoken language utterances and compare its results to those of an unsupervised topic model, which does not use labeled information.

Index Terms: spoken language understanding, generative models, gibbs sampling.

1. Introduction

In this paper, our goal is to extract meaning from natural language speech by learning to infer the speaker's intention. This information is then used to leverage understanding of the semantic structure of natural language utterances in spoken language understanding (SLU) systems. We demonstrate that extracting topical aspects of human utterances specifically benefits the domain detection problem of SLU. As a motivating example consider the following two human utterances: *"good places to eat"* and *"let's talk over lunch"*. The first refers to the *restaurant* domain, whereas the second can be classified into the *scheduling* or *calendar* domains (especially in personal assistance systems) since the action is related to the intent 'schedule a meeting'. Although one might argue that both utterances can be considered to be related to the restaurant domain, the latter one does not specifically indicate the aspect of the domain to which it relates. One potential challenge is that many of the important content-bearing words in the domain of interest may not be identified as part of the lexicon captured in training set. We show that extracting hidden topics in utterances can help to resolve the ambiguity in domain detection problems.

To understand natural language utterances, where there is very little evidence indicating the lexical context or domain, one should investigate long term dependencies between phrases and extract semantic concepts from previously seen utterances. To this end, various methods have been proposed for topic clustering [1, 5, 2, 3, 4, 6]. One approach that has become popular in SLU research is Latent Dirichlet Allocation (`LDA`) [7] because requires less supervision. Some of these related studies have implicitly used LDA models to learn underlying topic structure for purposes such as audio document clustering [10], speaker seperation [11], to name a few.

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 (and less dependent on the domain of the training set and the problems associated with human segmentation annotation). Thus, the aim of this study is to extract latent topic groupings from spoken language utterances (on the word or phrase level) using LDA, so that the utterances including the same latent topics (corresponding to hidden concepts) can be classified into same and/or similar semantic domains.

Despite the great success achieved with LDA, in this paper we raise two issues that are not commonly discussed. First, these models are generally built on documents represented as bags-of-words, meaning that multi-nomial topic distributions are defined for each document over ngrams extracted from the observed document sets. In other words, topics are sampled according to n-gram co occurrence statistics shared across documents. This raises an important issue related to building topic models on utterances. Compared to documents, utterances are relatively short, including one or at most two hidden topics; they add very little information to the word co-occurance statistics. In this paper, because we deal with the extraction of hidden concepts in utterances, we initially compile sets of in-domain utterances corresponding to documents and then build the topic models on these sets of utterances, instead of on single utterances. This prevents unsmoothed posterior latent topic distributions due to the sparsity of the bag of words in the utterance level models. We discuss the effects of this approach on the domain detection problem in the experiments.

The second non commonly-discussed challenge of using LDA for a specific recognition/classification task is that there is no guarantee that the latent topics learned will necessarily correspond to semantic classes, e.g., domains, that are previously defined. We introduce a new Semi Latent Dirichlet Allocation (`Semi-LDA`) based topic model specific to the utterance topic recognition task. In LDA models, the inference is based on sampling methods. In this paper we use Gibbs sampling for inference, a common implementation of Markov Chain Monte Carlo approximate inference methods for Bayesian inference. During Gibbs sampling we use an informative prior to determine the *latent topic-domain relations* in the training dataset. Our goal is to use the information that we learn from the in-domain training utterances and transfer onto individual test utterances, the domain information of which are unknown. Using learned topic-domain relationships, the model can predict the likelihood of a given utterance belonging to each possible domain class, as well as to the out-of-domain class.

We discuss leveraging one of the key semantic features of spoken and text utterances, namely the entities in topic models.

Since our approach implements approximate inference based on Gibbs sampling, contrary to feature based supervised classification methods, we utilize names in a different way. We compile a list of dictionaries specific to known domains and capture the entity names in utterances. For each new entity name, we introduce a pseudo-ngram to the vocabulary. In our experiments we found that using this semantic information as additional pseudo-words helped to improve the performance of the topic models.

In the rest of the paper we first tackle two major issues of unsupervised latent topic models and present results on the classification of utterances into semantic classes as follows: First we present our method for capturing latent topics that relate to our domains in §-2 and our new topic model that uses labeled information during extraction of latent topics from set of in-domain utterances in §-3. Later, we present an inference algorithm based on these clustering results in order to classify new utterances into one of the given semantic classes in §-4. § 5 is dedicated to experiments on real datasets to demonstrate the effects of our topic models and of lexicon extraction using named entities on the domain detection problem in comparison to standard LDA models.

2. Latent Topic Clustering

Assume we have seen a sequence of words/ngrams **w**=$\{w_1, w_2, .., w_n\}$ A topic model is a generative model that assumes a latent structure k comprising a set of words, **w**, and the concept used for the ith word, z_i, as an assignment of that word to one of the hidden topics. Below, we first describe the LDA model and later its extension to the utterance classification task specifically.

2.1. LDA Topic Modeling

In LDA the documents/utterances 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 utterance sequences are generated. For example, a typical utterance sequence can be composed of word n-grams like *"schedule"*, *"3pm"*, *"cafe plaza"*, etc, which may correspond to different semantic topics corresponding to aspecific domain[1]. Each word n-gram is represented as multinomial distribution over V words in the training data.

An utterance is generated by sampling a mixture of the semantic classes (topics) and then sampling n-grams conditioned on a particular semantic class. Each utterance is assumed to be drawn from a mixture of K shared topics, with topic z receiving a weight $\theta_z^{(u)}$ in utterance u. Each topic is a distribution over a shared vocabulary (lexicon) 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 $\phi^{(z)} \sim Dir(\beta), z = 1, .., K$.
2. For each N_u word n-grams $w_{u,n}$ in each utterance u:
   (a) Choose a topic $z_n \sim Mult(\theta^{(u_n)})$
   (b) Choose a word n-gram $w_n \sim \phi^{(z_n)}$

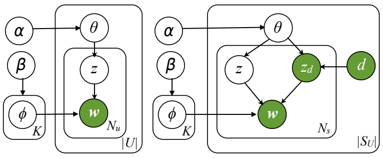

Figure 1: (*left*) Graphical model depiction of the LDA; (*right*) semi-latent domain topic model (Semi-LDA).

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 an utterance **u**=$w_1, .., w_{N_u}$ 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. The core equation of the LDA is the posterior probability of the topic label z_i for word i, conditioned on words 1 to n and all other topic labels 1 to n, given by

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

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

2.2. Semi Latent Topic Model - Semi-LDA

In the utterance classification task, we would like to attribute each utterance (in a given dialogue) to a possible semantic domain label. We also would like to build a more focused model, where there is a one-to-many map between the semantic domain classes and the latent topics. To enable this we use an informative prior during Gibbs sampling, which can utilize the word-domain frequency information from the training dataset. We implement the following approach for informative prior:

For those utterances for which we know the semantic class labels (training utterances), we sample from the topics designated for that semantic class. Similarly, for the unlabeled utterances whose semantic domain is not known, we sample topics from a list of possible semantic domains. At training time, we construct a lattice of n-gram frequencies per semantic domains to be used as prior information. During model training and inference, we use this lattice as restrictive information when generating each word in each utterance. Specifically, we reserve a list of latent topics z_d to sustain a correspondence between the latent topics and the semantic labels (classes). We also generate a number of other latent topics z_{ood} to be later labeled as *other* domains, i.e., for utterances that have lower posteriors for the rest of the labeled topics, z_d.

2.2.1. Pre-Processing for Latent Topic-Semantic Class Correspondence Discovery:

At training time, we are given a set of labeled utterances where the labels correspond to one of the semantic classes (domains), indicated by d in Fig. (1-*right*). As discussed in § 1, we build two different sets of models. One set of models is built on the individual utterance level, where each utterance is considered to be a document, as described in §2.1. In the second approach,

[1] In this study, the assumption is that each utterance is classified in a single semantic class.

we generate sets of utterances by randomly sampling from in-domain utterances. In the experiments, we keep the number of utterances to be sampled for each set as an input parameter. The underlying idea is to approximate the bag-of-words document structure, where there ismore evidence of semantic class, using utterances containing a few words, e.g., *"show me comedies playing downtown"*, which can only produce very sparse topic distributions and may result in weak assumptions for the domain of the utterance. The new generative model is described as follows:

A set of utterances S_U is a vector of N_s ngrams, $\mathbf{w}_s = \{w_{ns}\}_{n=1}^{N_s}$, where each $w_{ns} \in \{1, ..., V\}$, is chosen from a vocabulary of size V, and a vector of d domains, chosen from a set of semantic classes of size D. In addition, since we wish to discover templates from utterances that would allow attributions for bounded semantic concepts K in text, for a given set of utterances, the bag-of-ngrams are sampled from a list of possible domains, $d_{i=1,..D}$. The preprocessing steps for Semi-LDA are:

Step-1 Designate the first d topics to sample from known domains of the training dataset, leave the rest of the topics $K-d$ to the domains that are outside the defined semantic classes. Generate a lattice $\mathcal{L}_{w\times d}$, word frequencies by semantic domain based on the labeled training utterances.

Step-2: Build an Semi-LDA model on sets of in domain utterances S_U. This process is similar to the LDA model except that when sampling words for an utterance, whose domain is known a priori, we sample from the topics that are designated for that semantic class (domain). The generative process of Semi-LDA model (Fig. 1 *right*) can be formalized as:

1. Choose $\theta^{(s)} \sim Dir(\alpha)$, *s=1,..,*$|\mathcal{S}_\mathcal{U}|$, and choose $\phi^{(z)} \sim Dir(\beta)$, $z = 1, .., K$.
2. For each N_s word n-grams $w_{s,n}$ in each utterance $\mathcal{S}_\mathcal{U}$:
   (a) Find the possible domains $\tilde{d}_{w_{s,n}}$ for the $w_{s,n}$ based on the $\mathcal{L}_{w_{s,n}\times d}$ and later sample a topic $z_{d_n} \sim Mult(\theta^{(s_n)})$. If no possible domains present, sample a $z_{ood} \sim Mult(\theta^{(s_n)})$.
   (b) Choose a word n-gram $w_n \sim \phi^{(z_{d_n},\tilde{d}_{w,s_n})}$

A topic is sampled to generate each ngram using:

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

The indicator, $I[.]$, is used to eliminate those domains that the word n-gram $w_{s,n}$ has not been identified in the lattice $\mathcal{L}_{w_{s,n}\times d}$, 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 $w_{s,n}$ has been used in the restaurant and calendar domains, it is very likely that one of these domains will be chosen as the topic, z_d.

2.2.2. Labeling Latent Topics for Unlabeled Utterances

When sampling n-grams for labeled training utterances, we first sample from the possible topics z_d, which correspond to the semantic class of the utterance based on the lattice $\mathcal{L}_{w_{s,n}\times d}$. At testing time, we do not have the labeled utterances, thus we cannot use the informative prior in the same way as we did during model training. Instead, initially, we use the lattice structure from the training dataset to identify *possible* topics. In addition we let the algorithm sample from the out-of-domain topics, denoted as z_{ood}. Specifically, at testing time, using the uninformative prior of unsupervised LDA, we let the algorithm sample from both the out-of-domain topics, z's, as well as the possible domain-specific topics z_d when generating the words. This enables learning of the topic-semantic class relations $z_d - d$ as well as out of domain topics z_{ood} as shown in Fig. (1-*right*).

3. Inference for Domain Detection

At training, Semi-LDA enables sampling from topics designated as belonging to defined domains when generating words in utterances. From this process, we obtain the posterior latent topic-word distributions for in-domain topics, ϕ_{z_d} as well as out of domain topics $\phi_{z_{ood}}$. At testing time, we first predict the latent topic distributions over the words of each test utterance. Next, in order to predict the domain of a given test utterance, we execute an inference method akin to a language model, and calculate the domain likelihood of each utterance. Hence, we calculate a score corresponding to the likelihood of a test utterance given a domain as follows: The score of an utterance, $\mathbf{u_i}$=$w_1, .., w_{N_u}$, given a domain d is calculated by:

$$score(u_i|d) = p(z_d|\theta^{(d)}) \left(\prod_{n=1}^{N_u} p(w_n|z_d, \beta) \right) \quad (4)$$

Later, the best fitting (1-best) domain is determined by:

$$domain(u_i) = arg \max_d score(u_i|d) \quad (5)$$

4. Vocabulary Extension via Domain Dependent Entities

For text understanding, semantic features such as named entities play an important role for leveraging context information. For instance, the existence of a *"movie name"* entity and/or *"time information"* should improve the likelihood of the utterance being classified into the *"movie"* domain. In feature-based utterance classification models, such information is easily converted into additional features, binary or nominal, that would either indicate the existence of an entity in the utterance, or a set of features implying the existence of that particular entity.

We take a practical approach and annotate each utterance with a named entity using pre-compiled dictionaries. Based on a given domain, we compile domain-specific dictionaries including the names of movies, actors, movie directors, hotels, restaurants, cities, etc. Using the dictionaries, for a given utterance such as *"action movies directed by james cameron"*, we first capture that *"action"* is a movie-genre and add to the utterance a pseudo-ngram, $\langle movie - genre \rangle$, and similarly for *"James Cameron"* we append pseudo n-gram $\langle movie-director \rangle$. The new utterance then contains additional n-grams as: "$\langle movie - genre \rangle$ *action movies directed by* $\langle movie - director \rangle$ *james cameron*".

For topic models, we extend the vocabulary based on these pseudo-ngrams, which eanbles the use of named entity features during word generation through sampling latent topics. After we populate each utterance where we identify entities, we re-train the models with the new utterances using LDA and Semi-LDA and discuss the results in the experiments.

5. Experiments

Set up: Here, we assess the impact of semi-latent topic modeling when used for a domain detection task. Experiments were performed on the dataset obtained from the ASR output of an in-house data collection effort.

We use three different sets of document structures to build the semi latent topic models. First we train the models on individual utterences (ind); thus each utterance is considered to be an independent document. Later, we compile the documents using sets of utterances, $n = 10$ and use those sets as documents, rather than individual utterances, $(u10)$. We later construct one document per domain, compiled from all the utterances from a single domain. In the next subsection, we present the results of experiments on Semi-LDA models built based on these three sets of document structures, i.e,. Semi-LDA$_{(ind)}$, Semi-LDA$_{(u10)}$, Semi-LDA$_{(uall)}$. For fair comparison, as a baseline model, we used the unsupervised latent clustering model, standard LDA [7], which uses *uninformative* (random) prior for sampling topics for each n-gram. We build three different LDA models with the same three document sets used to build Semi-LDA models, i.e., LDA$_{(ind)}$, LDA$_{(u10)}$, LDA$_{(uall)}$.

We compile around 16K training utterances distributed among 25 domains, i.e., restaurants, movies, calendar, scheduling, transportation, weather, web, greeting, traffic, hotels, etc, which also include 'other' domain indicating the out of domain utterances. Out-of-domain utterances can be anything outside the 25 defined domains. Using the same dataset, we also experiment on a rather smaller domain set by merging utterances from domains containing inadequate data (e.g., ticket purchasing, traffic, weather, etc.) into a single domain designated as 'other'. We also compiled 1902 testing utterances and labeled them manually into 25 and 5 domains. The training and testing utterances were labeled manually by two annotators, where reliability rate was 70%.

Results and Discussions: We use the error rate of incorrect classification as the performance measure to compare different semantic domain classification models, which are summarized in Table-1. It can be observed that the Semi-LDA models can improve the semantic domain detection model performance in comparison to the rest of the models when there are more utterances in the documents (trained on sets of utterances). The best performance is achieved when the utterances are separated into individual domains and corresponding utterances are compiled together to form one document per domain, i.e., LDA$_{(uall)}$ and Semi-LDA$_{(uall)}$. This enables us to derive two conclusions: (*i*) topic models are the most powerful in capturing the hidden topic multi-nomials in relation to semantic units when the models are built on documents containing sentences/utterances with a satisfactory bag-of-words, (*ii*) using an informative prior in Gibbs sampling as in Semi-LDA, helps to explain the latent topic-domain relations much better than unsupervised LDA, which uses uninformative prior when generating words in documents.

Table 1: *Domain Detection Performance Measures in error.*

	25 Domain		5 Domain	
Model Name	Train	Test	Train	Test
LDA$_{(ind)}$	16.7%	18.4%	13.0%	13.4%
LDA$_{(u10)}$	14.2%	15.4%	10.9%	12.9%
LDA$_{(u_{all})}$	8.5%	13.7%	8.6%	12.2%
Semi-LDA$_{(ind)}$	12.2%	12.9%	12.2%	13.0%
Semi-LDA$_{(u10)}$	11.4%	11.8%	11.3%	12.4%
Semi-LDA$_{(u_{all})}$	**4.2%**	**9.7%**	**3.75%**	**9.93%**

Notice from Table-1 that when the domain size is increased, the error rate of the LDA models gets slightly worse, while that of the Semi-LDA model does not change as much. This indicates that the Semi-LDA is more robust in a multi-domain classification task when compared to LDA.

In Table-2 we show the experiment results, where the named entities are used as additional information, i.e., to extend the vocabulary used to build the topic models (as explained in §4). Using the new utterance sets, we build new LDA and Semi-LDA models. It is no surprise to us that all models perform better when additional information from named entities are utilized during lexicon construction. The conclusions we drive from the results in Table-1 applies here as well. The Semi-LDA model's error reduction in comparison to the LDA model's is statistically significant, as measured by a t-test.

Table 2: *Domain Detection Performance Measures using named entities in error.*

	25 Domain		5 Domain	
Model Name	Train	Test	Train	Test
LDA$_{(ind)}$	11.0%	15.4%	10.7%	12.0%
LDA$_{(u10)}$	10.9%	13.9%	6.4%	9.5%
LDA$_{(u_{all})}$	7.6%	11.9%	5.9%	8.3%
Semi-LDA$_{(ind)}$	3.7%	11.3%	2.4%	9.6%
Semi-LDA$_{(u10)}$	2.1%	8.7%	2.4%	8.5%
Semi-LDA$_{(u_{all})}$	**1.8%**	**8.14%**	**0.14%**	**6.7%**

6. Conclusion

In this paper, we analyze the effects of using labeled information in unsupervised latent topic clustering models on domain detection tasks. We present a semi-latent topic clustering model with a new informative prior that uses semantic class labels for utterances at training time. We show that integration of labeled information during Gibbs sampling significantly improves the the domain classification task performance when compared to unsupervised latent clustering models that use no-supervision at training time. We show that when the n-grams relating to real-world entities are populated with additional pseudo n-grams, we reduce the ambiguity of the semantic clusters and achieve further improvement in the domain detection performance.

The SemiLDA models can utilize unlabeled data which can help to reshape the prior information, as well generate the new lexicon. Recent work on domain detection [8, 9] indicate that exploiting query click logs has led to improvement on understanding user intent. In future work, we plan to use search queries during training to extract relationships within n-grams, which would eventually enable our model to handle natural language utterances of longer sequences.

7. References

[1] Chu-Carrol, J. and Carpenter, B., Vector-based natural language call routing, Computational Linguistics, 25(3) 361-358, 1999.

[2] Hoffman, T., Probabilistic latent semantic analysis. UAI, 1999.

[3] Tam, Y.C., and Schutz, T., Unsupervised language model adaptation using latent semantic marginals, Interspeeech, 2006.

[4] Hsu, B.J., and Glass, J., Style and topic language model adaptation using HMM-LDA. Interspeech 2004.

[5] Bellegarda, J. Large vocabulary speech recognition with multi-span statistical language models. IEEE Tran. on Speechand Audio Processing 8(1), 2000.

[6] Akita, Y. and Kawahara, T., Language model adaptation based on PLSA of topics and speakers. Interspeech 2004.

[7] Blei, D. and Ng, A.Y., Jordan, M.I., Latent Dirichlet Allocation, Machine Learning Research, 3:993-1022, 2003.

[8] Li, X., Wang, Y.-Y., Acero, A., Learning query intent from regularized click graphs, SIGIR 2008.

[9] Hakkani-Tur, D., Heck, L., Tur, G. Exploiting Query Click Logs For Utterance Domain Detection in Spoken Language Understanding, ICASSP 2011.

[10] Boulis, C., Ostendorf, M. Using symbolic prominence to help design feature subsets for topic classification and clustering of natural human-human conversations. Interspeech, 2005.

[11] Raj, B., Smaragdis, P. Shashanka, M.V.S., Latent Dicihlet decomposition for single channel speaker separation. ICASSP 2006.

[12] Iso, K.-I. Web-based topic language modeling for audio indexing, Proc. ICME'09 Proceedings of the IEEE Int. Conf. on Multimedia and Expo, 2009.