Exploiting Distance Based Similarity in Topic Models for User Intent Detection

Asli Celikyilmaz [1], Dilek Hakkani-Tur [1,2], Gokhan Tur [1,2], Ashley Fidler [1], Dustin Hillard [1]

[1] *Microsoft Speech Labs,* [2] *Microsoft Research*
Mountain View, CA, 94041
asli|dilek|gokhan.tur|aefidler|hillard@ieee.org

***Abstract*—One of the main components of spoken language understanding is intent detection, which allows user goals to be identified. A challenging sub-task of intent detection is the identification of intent bearing phrases from a limited amount of training data, while maintaining the ability to generalize well. We present a new probabilistic topic model for jointly identifying semantic intents and common phrases in spoken language utterances. Our model jointly learns a set of intent dependent phrases and captures semantic intent clusters as distributions over these phrases based on a distance dependent sampling method. This sampling method uses proximity of words utterances when assigning words to latent topics. We evaluate our method on labeled utterances and present several examples of discovered semantic units. We demonstrate that our model outperforms standard topic models based on bag-of-words assumption.**

I. Introduction

Spoken Language Understanding (SLU), a component of spoken dialogue systems, has been an active research area, in which the aim is to extract semantic meaning, such as speaker intention, from speech signals in order to provide natural human-to-machine or human-to-human interfaces [1]. Intent detection is the task of classifying natural language utterances into one or more previously defined semantic intent classes.

One of the challenges of semantic classifiers, such as those used to build intent determination models, is that they require operations that allow significant utterance variability. For instance, although the two utterances *'where is avatar playing'* and *'show me the nearest theatres in Mountain View'* are classified into the same semantic intent class, i.e., $find-theater$, there is no lexical overlap between them. Without an intent bearing n-gram determiner, the intent detection task becomes intrinsically challenging, not only because there are no a priori constraints on what the user might say, but also because the system must generalize from a tractably small amount of training data.

At first glance, this task can be solved using existing methods such as supervised methos (e.g., maximum entropy [2], or SVM [3]). A common approach is to build a multi-class classifier trained on the lexical and semantic features of the utterances [4], [5], [6], [7], [8]. While discriminitive approaches to semantic classification tasks have been shown to work well in intent determination on several domains, in practice, extracting intent bearing phrases based on the co-occurrence statistics of phrase patterns is challenging with these approaches, especially when the utterances in a given intent class are sparse.

The goal of our work is to provide a mechanism for understanding natural language utterances that goes beyond local and utterance level features. Specifically, we focus on unsupervised clustering that can capture latent topic clusters as distributions from a given document or paragraph. In recent work [9], [10], unsupervised latent variable models have been used to cluster utterances into semantic clusters using Bayesian inference, such Latent Dirichlet Allocation (LDA) [11]. 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.

While these methods have been shown to effectively extract semantic groupings of n-grams as latent topics and to improve the performance of SLU models, they are, in theory, bag-of-word models. Specifically, the inference is generally based on Gibbs sampling (a common implementation of Markov Chain Monte Carlo approximate inference methods for Bayesian inference). Hence, each word in the corpus is generated from a latent topic that has more emphasis on the relative lexical frequency of a word to each topic but less emphasis on the other words in the vicinity (the lexical context). Moreover LDA has an underlying exchangeability assumption, which refers to the invariance of a sequence of random variables to the permutations of their individual instances. Exchangeability is often considered an advantageous property, but a significant portion of the data in the text, image and audio domains is not exchangeable. Word sequences in language models, for example, are not exchangable because the relative position of each word in the sequence is important in this case. However, such restrictions are hard to model in topic models and often lead to approaches that are either specific to a certain modality or very complex.

Hence, our aim in this study is to extract latent topic groupings from spoken language utterances, in which the components, i.e., words or phrases, are non-exchangeable. In our models a word's or phrase's distributional similarity to another word or phrase is a function of the surrounding context each time it appears. This is similar to co-occurrence statistics between words. It measures to what extent the words/phrases can be similar given utterances in which they appear. For example, given that the sentence *'show me the*

978-1-4673-0367-5/11/$26.00 ©2011 IEEE

nearest ultra star cinemas' is labeled with the $find-theater$ intent, we would like our algorithm to have higher confidence in clustering *'amc cinemas in my neighborhood'* or *'regal cinemas nearby'* in one of the $find-theater$ clusters. This can be achieved by defining a distance function, which looks at lexical context so that the sampling of words like *'amc'*, *'regal'* or *'nearby'* will also be affected by the semantic class assignments of the words in the vicinity.

We present a new topic model, the *Distance Dependent Semi-Latent Topic Model (dd-SLDA)*, to capture latent topics from related utterances. Our algorithm extracts intent bearing constituents from given utterances based on a distance function between a current word and the words in its vicinity. Because we use intent labeled utterances to assign each semantic cluster to one of the set of predefined intent clusters, labeling a new utterance with an intent cluster is straightforward and does not require additional classification methods. We also reserve a small number of clusters to capture out-of-intent clusters. In latent topic models there is usually no guarantee that the latent topics learned will necessarily correspond to defined semantic intent classes. To resolve this issue at Gibbs sampling time of dd-SLDA, we use an informative prior to determine the latent topic-intent relations, thereby constraining the word-topic assignments in addition to the vicinity of words defined via a distance method.

Despite the great success achieved with the topic models such as LDA [11], in this paper we raise another issue that is not commonly discussed. These models are generally built on documents, meaning that multinominal topic distributions are defined for each document over n-grams extracted from complete set of observed documents. This raises an important issue related to building topic models for 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-occurrence statistics. Because we deal with the extraction of hidden concepts in utterances, to be as close to document representation in topic models, we initially compile sets of utterances with similar semantic intents, and then build the distance dependent topic models on these sets of utterances, instead of on single utterances (similar to [9]). This prevents unsmoothed posterior latent topic distributions due to the sparsity of the n-grams in the utterance level models. We discuss the effects of this approach on the intent detection problem in the experiments.

In the rest of this paper we first tackle two major issues of unsupervised latent topic models and then present results on the classification of utterances into semantic intent classes. The structure is as follows: First we present the standard LDA model in section-2 and the distance dependent sampling method for capturing latent topics related to our semantic intents in section-3. Also in section-3 we present how our topic model uses labeled information during the extraction of latent topics from a set of similar intent utterances. In section-4, we present an inference algorithm based on these clustering results, in order to classify new utterances into one of a set of given semantic classes. In section-5 we present the results of experiments on real datasets to demonstrate the effects of our topic models in comparison to standard LDA models.

II. Latent Dirichlet Allocation - LDA

A topic model is a generative model that assumes a latent structure k comprising a set of words, $\mathbf{w}$, and the concept used for the mth word, z_m, as an assignment of that word to one of the hidden topics. We start at the level of the observed frequencies of n-grams given intent clusters from labelled data, and then work our way up to the distance dependent sampling algorithm. As is customary in topic model learning applied to text data, we divide each sentence into bag-of-n-grams (up to three grams) represented as $w_m = 1, \ldots, V$, where V is the vocabulary size and assume we have already seen a sequence of words/ngrams $\mathbf{w}=\{w_1, w_2, .., w_n\}$.

In LDA the documents/utterances are modeled as distributions over sets of hidden topics and each hidden topic is 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 can be composed of word n-grams, which may represent specific intent, such as *'lunch'*, *'3pm'*, *'cafe plaza'*, etc.

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 V 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 can be formalized as follows:

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 $w_n \sim \phi^{(z_n)}$

Here α and β are fixed hyper-parameters; we need to estimate parameters θ for each document and ϕ for each topic. From the expectation of Dirichlet distributions, the probability of an utterance $\mathbf{u}=w_1, .., w_{N_u}$ 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, 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

$$P(z_m|z_{n\setminus m}, w_n) \propto \frac{n_{z_m,n\setminus m}^{(w_m)} + \beta}{n_{z_m,n\setminus m}^{(.)} + W\beta} \cdot \frac{n_{z_m,n\setminus m}^{(u_m)} + \alpha}{n_{.,n\setminus m}^{(u_m)} + K\alpha} \quad (2)$$

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

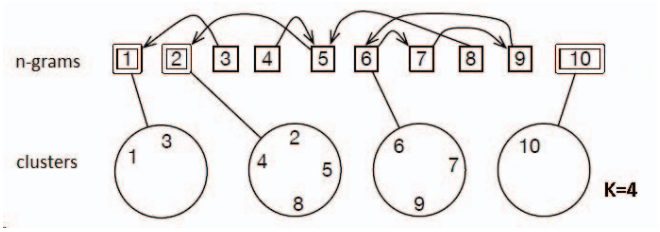

Fig. 1. An illustration of a distance based sampling method. The process operates at the level of n-gram assignment where each n-gram chooses another n-gram (indicated by **arrows**) or sampled individually according a user defined distance function. The **squares** represent the n-grams and the **double lines** around the squares represent words that are indicative of the intent clusters (circles) obtained from labeled training data. (The **lines** between words denoted as double-lined squares and circles (intent clusters) indicate that such words are indicative of one of the intents extracted from the labeled utterances.) An n-gram is assigned to the same cluster as another already sampled n-gram (e.g. 1-3, 2-4, 9-6, etc.) if their proximity, as measured by the distance function is large. If the n-gram is not assigned to any of the predefined clusters, it is assigned to one of the out-of-intent clusters (the cluster on the right). Thus, the cluster assignments are derived from the n-grams and their vicinity terms in the given sequence (utterances or documents).

III. Distance Dependent Topic Model

Distance dependent topic models [12] expand the palette of topic models by representing the partition of the n-gram assignments rather than topic assignments. While the traditional topic model connects words to topics, the distance dependent topic models connect words to other words. The distance dependent topic model presented in this paper is a specification of the standard LDA models as well as Distance Dependent Chinese Restaurant Processes (CRP) [13], a class of non-parametric Bayesian models which assigns customers to customers instead of customers to tables in the standard CRP. Specifically, the CRP defines a distribution over partitions that embodies the assumed prior distribution over cluster structures [13]. A Chinese restaurant with an infinite number of tables is considered to employ a sequential process by which customers enter the restaurant and each sit down at a randomly chosen table. After N customers have sat down, their configuration at the tables represents a random partition.

In this paper we present a specification of the distance dependent CRP, where the number of tables is fixed (similar to standard LDA models) to K topics. We introduce two new additions to the sampling algorithm of our Distance Dependent Semi-Latent Topic Models (`dd-SLDA`). First, we extend the gibbs sampling method by introducing a distance dependent sampling algorithm for cluster assignments, which is important to capture rare constituents in utterances indicating specific intents. Specifically, with the `dd-SLDA` we define a user-specific function for sampling n-grams given other n-grams in an utterance or sequence of utterances such as a dialogue. Thus, the n-gram assignment to clusters should also depend on the distance between the word and the other words in its immediate vicinity (see Fig. 1).

Secondly, since our task is the classification of utterances into user defined semantic intents, we would like to attribute each utterance to a possible semantic intent label. We do this using a more focused model, where there is a one-to-many map between the semantic intent classes and the latent topics, namely, we have semi-latent topics corresponding to each semantic intent. To enable this, we enrich the Gibbs sampling with an informative prior, which can utilize the word-domain frequency information from the training dataset. In the next section, we present the details of these two new extensions of the sampling methods and their implementation to the overall `dd-SLDA` model.

A. Distance Function for Gibbs Sampling

The main difference between distance dependent clustering and standard LDA is that in distance based sampling, the n-grams are sampled to clusters together with other n-grams, instead of being individually sampled directly to clusters. Connected groups of n-grams are only implicitly assigned together to the same cluster. In Fig. 1 we illustrate the distance based sampling method with four clusters and ten words.

Given a set of utterances, the *i*th n-gram word is assigned to some cluster either together with another n-gram j (denoted as $c_i = j$) with a probability proportional to a decreasing function of the distance between the two: $f(d_{ij})$, or alone if the similarity is less than a threshold, α. Hence the larger the distance, the less likely a word is to be sampled from the same latent topic cluster with some other n-gram. This leads to the following multinomial distribution over n-gram assigments conditioned on distances $D \in \Re^{N \times N}$, where N is the number of n-grams and the decay function $f : \Re^{+} \rightarrow \Re^{+}$ needs to be non-increasing and have $f(\infty) = 0$,

$$p(c_i = j | d, \alpha) = \begin{cases} f(d_{ij}) & i \neq j \\ \alpha & i = j \end{cases} \quad (3)$$

We have experimented with several distance functions mainly based on the vicinity information. We want the distance function to be parameterized (e.g., in the case of exponential decay), thus we used the exponential distance method with the parameter v indicates the number of words that the linking word are apart.We used up to two words as the window for the vicinity measure.

$$f_v(d_{ij}) = exp(-d_{ij}/v) \quad (4)$$

where distance d is defined as the proximity of the words (measured by the number of words between the two words) given that utterance u, normalized by the frequency of the two

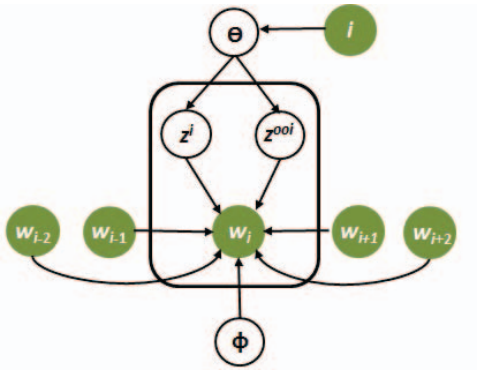

Fig. 2. Graphical Representation of the Distance Dependent Semi-Latent Topic Model- *dd*-SLDA. The filled circles indicate observed variables (n-grams in sets of utterances represented as a plate) and arrows indicate the conditional variables in the directed acyclic graph. Each word w_j is sampled from either one of the latent semantic intent classes z^i or one of the out-of-intent classes z^{ooi} conditioned on the distance between the vicinity of words (represented by the conditional distribution - the arrows between w_j and the $w_{j\neq j-t}$),$t = 1, 2...$ as well as on the prior information of the observed frequency of the given word w_j in utterances with same intents, indicated by the observed variable i. The θ and ϕ are the prior Dirichlet distributions that the sets of utterance-topic and topic-n-grams (w) are sampled from. Hyperparameters are not shown.

words appearing in the corpus together given the proximity constraint:

$$d_{ij}|u = \frac{prox(i,j)_u}{\sum_u^U prox(i,j)_u} \tag{5}$$

At training time, the distance based probablity in Eq. (3) is used as an additional constraint in sampling n-grams to latent intent or out-of-intent clusters.

B. Pre-Labeling Latent Clusters

For those n-grams in 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 intent is not known, we sample topics from a list of possible semantic intents. At training time, we construct a lattice of n-gram frequencies per semantic intents 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_i to sustain a correspondence between the latent topics and the semantic intent labels (classes). We also generate a number of other latent topics z_{ooi} to be later labeled as *other* out-of-intent clusters, i.e., for utterances that have lower posteriors for the rest of the labeled topics, z_i.

C. `dd-SLDA` Topic Model

We are given a set of labeled utterances where the labels correspond to one of the user defined intents, indicated by i in Fig. (2). As discussed in section 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. In the second approach, we generate sets of utterances by randomly sampling from utterances of similar intents (similar to [9]). 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 as close to a document structure as possible, where there is more evidence of semantic class. Using utterances that contain only a few words, e.g., *"show me comedies playing downtown"* can only produce very sparse topic distributions and may result in weak assumptions for the intent of the utterance; whereas a list of utterances with the same intent will contain more n-grams that intent bearing phrases can be easily extracted. 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 i intents, chosen from a set of semantic classes of size I. In addition, since we wish to discover templates from utterances that would allow attributions for bounded semantic concepts K, for a given set of utterances, each n-gram is sampled from a list of possible intents. The preprocessing steps for `dd-SLDA` are:

Step-1 Designate the first i topics to sample from known intents of the training dataset, leave the rest of the topics $K-i$ to the intents that are outside the defined semantic classes. Generate a lattice $\mathcal{L}_{w\times i}$, of word frequencies by semantic intent based on the labeled training utterances. Also keep the word sequence information given the sets of utterances for distance based sampling.

Step-2: Build a `dd-SLDA` model on sets of utterances S_I with the same intent. This process is similar to the `LDA` model except that when sampling words for an utterance, whose intent is known a priori, we sample from the topics that are designated for that semantic classes (intents). The generative process of `dd-SLDA` model (Fig. 2) 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) Define a decay function f, sequential distance $d_{j,j-t}$, $t = 1, 2$ for two n-grams, the threshold for out-of-intent cluster sampling.
3) For each N_s word n-grams $w_{s,n}$ in each utterance $\mathcal{S}_\mathcal{U}$:
   a) Find the possible intents $\tilde{i}_{w_{s,n}}$ for the $w_{s,n}$ based on the $\mathcal{L}_{w_{s,n}\times i}$,
   b) Find the nearest T n-grams based on distance d function,
   c) Sample a topic $z_{i_n} \sim Mult(\theta^{(s_n)})$ using Eq.(6). If no possible intents present, sample a $z_{ooi} \sim Mult(\theta^{(s_n)})$.
   d) Choose a word n-gram $w_n \sim \phi^{(z_{i_n}, \tilde{i}_{w,s_n}, d(j,j-t|t=1,2))}$

A topic is sampled to generate each ngram using:

$$p(z = k|w_n, i, \mathbf{z}_{-i}, d(j, j-t), t = 1, 2) = \arg\max_t P(z_k|z_{\mathbf{i}\setminus\mathbf{i}}, w_n) * I[w_{s,n} \in \tilde{i}_{w_{s,n}}] * p(c_j = j - t|d, \alpha) \tag{6}$$

Here distance method is based on the immediate words in the vicinity that are at most two words apart, i.e., $t = 1, 2$. The

indicator, $I[.]$, is used to eliminate those intents where the word n-gram $w_{s,n}$ has not been identified in the lattice $\mathcal{L}_{w_{s,n}\times i}$, hence the designated topics are not sampled from them. The last term constrains the assignment of the n-gram w_n to one of the intent clusters within the vicinity that the t vicinity words have been sampled from. The $argmax$ enables sampling the nth word based on the nearest words in proximity (choses the the topic that the most closest n-gram have sampled from as well as the prior frequency of the word given that topic). Instead of using random topic sampling, e.g., an uninformative prior of unsupervised LDA, we use an informative prior as explained above 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 $find-movie$ and $find-theater$ intents, and its nearest neighbors are also sampled from these intents, it is very likely that one of these intent clusters will be chosen as the topic, z_i.

D. Labeling Latent Topics for Unlabeled Utterances

In our previous work [9]. we have generated a simple latent topic labeling method. Here we extend our approach by considering the word pair information as follows: When sampling n-grams for labeled training utterances, we first sample from the possible topics z_i, which correspond to the semantic class of the utterance based on the lattice $\mathcal{L}_{w_{s,n}\times i}$ as well as the probability of words in the vicinity. 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, we use the lattice structure from the training dataset to identify *possible* topics. In addition we let the algorithm sample from the out-of-intent topics, denoted as z_{ooi}. Specifically, at testing time, using the uninformative prior of unsupervised LDA, we let the algorithm sample from both the out-of-intent topics, z_{ooi}'s, as well as the possible intent-specific topics z_i together with the topic information of the proximate words when generating any word.

IV. Inference for Intent Detection

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

$$score(u_m|i) = p(z_i|\theta^{(i)}) \left(\prod_{n=1}^{N_u} p(w_n|z_d, \beta) \right) \quad (6)$$

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

$$domain(u_m) = arg\max_i score(u_m|i) \quad (7)$$

Fig. 3. Example utterances grouped by intent (colored blue). Intent bearing phrases and words related within proximity are colored red.

V. Experiments

In this section we describe in detail our data set and present experiments and their results.

Data Set: Our data set consists of spoken language utterances in the movies domain. Each utterance is labeled with different semantic intents, e.g., $find$-$movie$, get-$trailer$, $rate$-$review$-$compare$-$movies$, etc. There is also an *other* intent which indicates that an intent is not covered. There are 4200 utterances in total, collected from several sources. Examples of utterances by intent are shown in Fig. 3. Training and testing utterances were labeled manually by two annotators, where the inter-annotator aggreement as measured by Kappa was 80%.

Task: We perform experiments using two different sets of document structures. First we train our topic models on individual utterances, taking utterances as documents in the topic models. Second, we group utterances into sets of 10 and treat those sets as documents, rahter than individual utterances. We later construct one document per intent, compiled from all the utterrences from a single intent.

Results: We use the error rate of incorrect classification as the performance measure to compare different semantic intent classication models, which is summarized in Table 1. Note

TABLE I
Intent detection accuracy of our model to the standard LDA trained on (1) single utterances, (2) sets of utterances selected from intent-specific utterances ($n = 10$) (3) documents compiled from all utterances per intent

Model	Test set Accuracy
LDA (1)	0.65
`dd-SLDA` (1)	0.70
LDA (2)	0.68
`dd-SLDA` (2)	**0.84**
LDA (3)	0.75
`dd-SLDA` (3)	**0.88**

that, when topic models are build on individual utterances the intent detection performance is significantly lower compared to the topic models trained on a larger set of utterances. This is mainly due to the fact that although the utterances are natural language, they tend to be short comprared the the documents or paragraphs which are usually used to train the topic models.

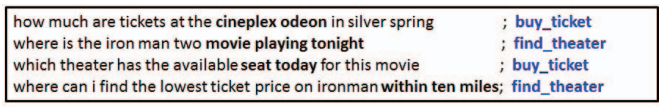

Fig. 4. Examples of utterrances correctly labeled by our system but not correctly labeled by the standard LDA baseline. The key intent bearing n-grams are hightlighted. Notice that the words are not the most common intent bearing words, therefore it is difficult for the samplers that do not consider long term dependencies given utterances.

The sparse distributions of word-topics and topic-utterenaces in these models hurt the posterior likelihood of words/n-grams.

On the other hand, our models which use a new Gibbs sampling with an informative prior outperforms the uninformative prior of the standard LDA model. The most gain is obtained when our models are trained on sets of utterances where the intent bearing distuributions, a.k.a., latent semantic intent clusters, are approximated in relation to other intents. The new distance based sampling enables sampling of words that are indicative of an intent along with the other words that are indicative of the same intent.

Discussion on Extraction of Intent Indicator Phrases: In Fig. 4 we show some examples of utterrances correctly labeled by our system but not correctly labeled by the standard LDA baseline. Our algorithm can capture the key intent bearing n-grams based on the distance dependent sampling algorithm. Such n-grams does not need to be the most frequent word given an intent. The distance based sampling enables samplings words not only based on co-occurance statatiscs given a cluster, but also based on word pair information. This menthod enables two word pairs (that are not necessarily frequently observed in the corpus) ending in the same intent cluster with high probability to the same intent cluster.

VI. Conclusion

We have presented a probabilistic topic model for identifying hidden semantic intent classes and intent bearing constituents from spoken language utterances. The model is relatively simple and admits an efficient Gibbs sampling inference procedure which enables long term dependencies. We have demonstrated on evaluation task that our model outperform an applicable baseline by a considerable margin.

As a future work, we plan to use unlabeled utterances with no prior information to extend our vocabulary given the domain of interest. We will generate a boosting type learning algorithm, where utterances with words that are found to have high degree of confidence in semantic intent classes based on our method, will be labeled with the corresponding intents. With this iterative method, our models can generalize well to unseen data.

Acknowledgements

Many thanks for Ye-Yi Wang and Xiao Li from Microsoft Research for useful discussions and their input. We would also like to thanks the anonymous reviewers for their valuable insignts and directions.

References

[1] G. Tur and R. D. Mori, Eds., *Spoken Language Understanding : Systems for extracting semantic information from speech (eds)*. John Wiley and Sons, 2011.
[2] A. Berger, S. Pietra, and V. Pietra, "A maximum entropy approach to natural language processing," *Journal*, vol. 22(1), pp. 39–71, 1996.
[3] V. Vapnik, Ed., *Statistical Learning Theory*. Springer., 1995.
[4] P. Haffner, G. Tur, and J. Wright, "Optimizing svms for complex call classification," *Int. Conference on Acustics, Speech and Signal Processing*, 2003.
[5] S. Yaman, L. Deng, D. Yu, Y. Wang, and A. Acero, "An integrative and discriminative technique for spoken utterance classification," *IEEE Trans. on Audio, Speech and Language Processing*, 2008.
[6] X. Li, Y. Wang, and A. Acero, "Learning query intent from regularized click graphs," *SIGIR*, 2008.
[7] M. Saraclar and B. Roark, "Joint discriminitive language modeling and utterance classification," *Int. Conference on Acoustics, Speech and Signal Processing*, 2005.
[8] M. Karahan, D. Hakkani-Tur, G. Riccardi, and G. Tur, "Combining classifiers for spoken language understanding," *IEEE ASRU*, 2003.
[9] A. Celikyilmaz, D. Hakkani-Tur, and G. Tur, "Multi-domain spoken language understanding with approximate inference," *InterSpeech*, 2011.
[10] M. Dowman, V. Savova, T. Griffiths, K. Kording, J. Tenenbaum, and M. Pruver, "A probabilistic model of meetings that combines words and discourse features." 2008.
[11] D. M. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," 2003.
[12] D. M. Blei and P. Fraizer, "Distnace dependent chinese restaurant process," 2010.
[13] J. Pitman, "Combinatorial stochastic processes," *Lecture Notes on St. Flour Summer School, NY*, 2002.