A Joint Model for Discovery of Aspects in Utterances

Asli Celikyilmaz
Microsoft
Mountain View, CA, USA
asli@ieee.org

Dilek Hakkani-Tur
Microsoft
Mountain View, CA, USA
dilek@ieee.org

Abstract

We describe a joint model for understanding user actions in natural language utterances. Our multi-layer generative approach uses both labeled and unlabeled utterances to jointly learn aspects regarding utterance's target domain (e.g. movies), intention (e.g., finding a movie) along with other semantic units (e.g., movie name). We inject information extracted from unstructured web search query logs as prior information to enhance the generative process of the natural language utterance understanding model. Using utterances from five domains, our approach shows up to 4.5% improvement on domain and dialog act performance over cascaded approach in which each semantic component is learned sequentially and a supervised joint learning model (which requires fully labeled data).

1 Introduction

Virtual personal assistance (VPA) is a human to machine dialog system, which is designed to perform tasks such as making reservations at restaurants, checking flight statuses, or planning weekend activities. A typical spoken language understanding (SLU) module of a VPA (Bangalore, 2006; Tur and Mori, 2011) defines a structured representation for utterances, in which the constituents correspond to meaning representations in terms of slot/value pairs (see Table 1). While target *domain* corresponds to the context of an utterance in a dialog, the dialog act represents overall intent of an utterance. The *slots* are entities, which are semantic constituents at the word or phrase level. Learning each component

Sample utterances on 'plan a night out' scenario

(I) Show me theaters in [*Austin*] playing [*iron man 2*].
(II)I'm in the mood for [*indian*] food tonight, show me the ones [*within 5 miles*] that have [*patios*].

Extracted Class and Labels

Domain	Dialog Act	Slots=Values
(I) Movie	find theater	Location=*Austin* Movie-Name= *iron man 2*
(II) Restaurant	find restaurant	Rest-Cusine=*indian* Location=*within 5 miles* Rest-Amenities= *patios*

Table 1: Examples of utterances with corresponding semantic components, i.e., domain, dialog act, and slots.

is a challenging task not only because there are no *a priori* constraints on what a user might say, but also systems must generalize from a tractably small amount of labeled training data. In this paper, we argue that each of these components are interdependent and should be modeled simultaneously. We build a joint understanding framework and introduce a multi-layer context model for semantic representation of utterances of multiple domains.

Although different strategies can be applied, typically a cascaded approach is used where each semantic component is modeled separately/sequentially (Begeja et al., 2004), focusing less on interrelated aspects, i.e., dialog's domain, user's intentions, and semantic tags that can be shared across domains. Recent work on SLU (Jeong and Lee, 2008; Wang, 2010) presents joint modeling of two components, i.e., the domain and slot or dialog act and slot components together. Furthermore, most of these systems rely on labeled training utterances, focusing little on issues such as information sharing between the discourse and word level components across different domains, or variations in use of language. To deal with de-

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 330–338,
Jeju, Republic of Korea, 8-14 July 2012.

pendency and language variability issues, a model that considers dependencies between semantic components and utilizes information from large bodies of unlabeled text can be beneficial for SLU.

In this paper, we present a novel generative Bayesian model that learns domain/dialog-act/slot semantic components as latent aspects of text utterances. Our approach can identify these semantic components simultaneously in a hierarchical framework that enables the learning of dependencies. We incorporate prior knowledge that we observe in web search query logs as constraints on these latent aspects. Our model can discover associations between words within a multi-layered aspect model, in which some words are indicative of higher layer (meta) aspects (domain or dialog act components), while others are indicative of lower layer specific entities.

The contributions of this paper are as follows:
(***i***) construction of a novel Bayesian framework for semantic parsing of natural language (NL) utterances in a unifying framework in §4,
(***ii***) representation of seed labeled data and information from web queries as informative prior to design a novel utterance understanding model in §3 & §4,
(***iii***) comparison of our results to supervised sequential and joint learning methods on NL utterances in §5. We conclude that our generative model achieves noticeable improvement compared to discriminative models when labeled data is scarce.

2 Background

Language understanding has been well studied in the context of question/answering (Harabagiu and Hickl, 2006; Liang et al., 2011), entailment (Sammons et al., 2010), summarization (Hovy et al., 2005; Daumé-III and Marcu, 2006), spoken language understanding (Tur and Mori, 2011; Dinarelli et al., 2009), query understanding (Popescu et al., 2010; Li, 2010; Reisinger and Pasca, 2011), etc. However data sources in VPA systems pose new challenges, such as variability and ambiguities in natural language, or short utterances that rarely contain contextual information, etc. Thus, SLU plays an important role in allowing any sophisticated spoken dialog system (e.g., DARPA Calo (Berry et al., 2011), Siri, etc.) to take the correct machine actions.

A common approach to building SLU framework is to model its semantic components separately, assuming that the context (domain) is given *a priori*. Earlier work takes dialog act identification as a classification task to capture the user's intentions (Margolis et al., 2010) and slot filling as a sequence learning task specific to a given domain class (Wang et al., 2009; Li, 2010). Since these tasks are considered as a pipeline, the errors of each component are transfered to the next, causing robustness issues. Ideally, these components should be modeled simultaneously considering the dependencies between them. For example, in a local domain application, users may require information about a sub-domain (*movies, hotels, etc.*), and for each sub-domain, they may want to take different actions (*find* a movie, *call* a restaurant or *book* a hotel) using domain specific attributes (e.g., *cuisine type* of a restaurant, *titles* for movies or *star-rating* of a hotel). There's been little attention in the literature on modeling the dependencies of SLU's correlated structures.

Only recent research has focused on the joint modeling of SLU (Jeong and Lee, 2008; Wang, 2010) taking into account the dependencies at learning time. In (Jeong and Lee, 2008), a triangular chain conditional random fields (Tri-CRF) approach is presented to model two of the SLU's components in a single-pass. Their discriminative approach represents semantic slots and discourse-level utterance labels (domain or dialog act) in a single structure to encode dependencies. However, their model requires fully labeled utterances for training, which can be time consuming and expensive to generate for dynamic systems. Also, they can only learn dependencies between two components simultaneously.

Our approach differs from the earlier work- in that- we take the utterance understanding as a multi-layered learning problem, and build a hierarchical clustering model. Our joint model can discover domain $\boldsymbol{D}$, and user's act $\boldsymbol{A}$ as higher layer latent concepts of utterances in relation to lower layer latent semantic topics (slots) $\boldsymbol{S}$ such as named-entities (*"New York"*) or context bearing non-named entities (*"vegan"*). Our work resembles the earlier work of PAM models (Mimno et al., 2007), i.e., directed acyclic graphs representing mixtures of hierarchical topic structures, where upper level topics are multinomial over lower level topics in a hierarchy. In an analogical way to earlier work, the $\boldsymbol{D}$ and $\boldsymbol{A}$ in our

approach represent common co-occurrence patterns (dependencies) between semantic tags ***S*** (Fig. 2). Concretely, correlated topics eliminate assignment of semantic tags to segments in an utterance that belong to other domains, e.g., we can discover that *"Show me vegan restaurants in San Francisco"* has a low probably of outputting a *movie-actor* slot. Being generative, our model can incorporate unlabeled utterances and encode prior information of concepts.

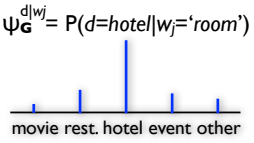
3 Data and Approach Overview

Here we define several abstractions of our joint model as depicted in Fig. 1. Our corpus mainly contains NL utterances ("*show me the nearest dimsum places*") and some keyword queries ("*iron man 2 trailers*"). We represent each utterance *u* as a vector $\mathbf{w_u}$ of N_u word n-grams (segments), w_{uj}, each of which are chosen from a vocabulary W of fixed-size *V*. We use entity lists obtained from web sources (explained next) to identify segments in the corpus. Our corpus contains utterances from K_D=4 main **domains**:$\in$ {*movies, hotels, restaurants, events*}, as well as out-of-domain *other* class. Each utterance has one **dialog act** (*A*) associated with it. We assume a fixed number of possible dialog acts K_A for each domain. Semantic Tags, **slots** (*S*) are lexical units (segments) of an utterance, which we classify into two types: domain-independent slots that are shared across all domains, (*e.g., location, time, year, etc.*), and domain-dependent slots, (*e.g. movie-name, actor-name, restaurant-name, etc.*). For tractability, we consider a fixed number of latent slot types K_S. Our algorithm assigns domain/dialog-act/slot labels to each topic at each layer in the hierarchy using labeled data (explained in §4.)

We represent domain and dialog act components as meta-variables of utterances. This is similar to author-topic models (Rosen-Zvi et al., 2004), that capture author-topic relations across documents. In that case, words are generated by first selecting an author uniformly from an observed author list and then selecting a topic from a distribution over words that is specific to that author. In our model, each utterance u is associated with domain and dialog act topics. A word w_{uj} in *u* is generated by first selecting a domain and an act topic and then slot topic over words of u. The domain-dependent slots in utterances are usually not dependent on the dialog act. For instance, while *"find* [*hugo*] *trailer"* and *"show me where* [*hugo*] *is playing"* have both a movie-name slot ("*hugo*"), they have different dialog acts, i.e., find-trailer and find-movie, respectively. We predict posterior probabilities for domain $\tilde{P}(d \in D|u)$ dialog act $\tilde{P}(a \in A|ud)$ and slots $\tilde{P}(s_j \in S|w_{uj}, d, s_{j-1})$ of words w_{uj} in sequence.

To handle language variability, and hence discover correlation between hierarchical aspects of utterances[1], we extract prior information from two web resources as follows:

Web n-Grams (*G*). Large-scale engines such as Bing or Google log more than 100M search queries each day. Each query in the search logs has an associated set of URLs that were clicked after users entered a given query. The click information can be used to infer domain class labels, and therefore, can provide (noisy) supervision in training domain classifiers. For example, two queries ("*cheap hotels Las Vegas*" and "*wine resorts in Napa*"), which resulted in clicks on the same base URL (e.g., www.hotels.com) probably belong to the same domain ("hotels" in this case).



Given query logs, we compile sets of in-domain queries based on their base URLs[2]. Then, for each vocabulary item $w_j \in W$ in our corpus, we calculate frequency of w_j in each set of in-domain queries and represent each word (e.g., "room") as a discrete normalized probability distribution $\psi^j_{\mathbf{G}}$ over K_D domains $\{\psi^{d|j}_{\mathbf{G}}\} \in \psi^j_{\mathbf{G}}$. We inject them as nonuniform priors over domain and dialog act parameters in §4.

Entity Lists (*E*). We limit our model to a set of *named-entity* slots (*e.g., movie-name, restaurant-name*) and *non-named entity* slots (e.g., *restaurant-cuisine*, *hotel-rating*). For each entity slot, we extract a large collection of entity lists through the url's on the web that correspond to our domains, such as movie-names listed on IMDB, restaurant-names on OpenTable, or hotel-ratings on tripadvisor.com.

[1]Two utterances can be intrinsically related but contain no common terms, e.g., *"has open bar"* and *"serves free drinks"*.

[2]We focus on domain specific search engines such as IMDB.com, RottenTomatoes.com for movies, Hotels.com and Expedia.com for hotels, etc.

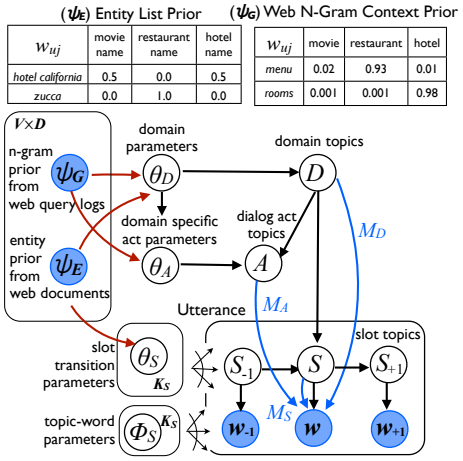


Figure 1: Graphical model depiction of the MCM. $\mathbf{D}, \mathbf{A}, \mathbf{S}$ are domain, dialog act and slot in a hierarchy, each consisting of K_D, K_A, K_S components. Shaded nodes indicate observed variables. Hyper-parameters are omitted. Sample informative priors over latent topics ψ_E and ψ_G are shown. Blue arrows indicate frequency of vocabulary terms sampled for each topic.

We represent each entity list as observed nonuniform priors ψ_E and inject them into our joint learning process as V sparse multinomial distributions over latent topics D, and S to "guide" the generation of utterances (Fig. 1 top-left table), explained in §4.

4 Multi-Layer Context Model - MCM

The generative process of our multi-layer context model (MCM) (Fig. 1) is shown in Algorithm 1. Each utterance u is associated with $d = 1..K_D$ multinomial domain-topic distributions θ^d_D. Each domain d, is represented as a distribution over $a = 1,..,K_A$ dialog acts θ^{da}_A ($\theta^d_D \rightarrow \theta^{da}_A$). In our MCM model, we assume that each utterance is represented as a hidden Markov model with K_S slot states. Each state generates n-grams according to a multinomial n-gram distribution. Once domain D_u and act A_{ud} topics are sampled for u, a slot state topic S_{ujd} is drawn to generate each segment w_{uj} of u by considering the word-tag sequence frequencies based on a simple HMM assumption, similar to the content models of (Sauper et al., 2011). Initial and transition probability distributions over the HMM states are sampled from Dirichlet distribution over slots θ^{ds}_S. Each slot state s generates words according to multinomial word distribution ϕ^s_S. We also keep track of the frequency of vocabulary terms w_j's in a $V \times K_D$ matrix M_D. Every time a w_j is sampled for a domain d, we increment its count, a degree of domain bearing words. Similarly, we keep track of dialog act and slot bearing words in $V \times K_A$ and $V \times K_S$ matrices, M_A and M_S (shown as red arrows in Fig 1). Being Bayesian, each distribution θ^d_D, θ^{ad}_A, and θ^{ds}_S is sampled from a Dirichlet prior distribution with different parameters, described next.

Algorithm 1 Multi-Layer Context Model Generation

for each domain $d \leftarrow 1, ..., K_D$
  draw domain dist. $\theta^d_D \sim Dir(\alpha^\star_D)^\dagger$,
  for each dialog-act $a \leftarrow 1, ..., K_A$
    draw dialog act dist. $\theta^{da}_A \sim Dir(\alpha^\star_A)$,
  for each slot type $s \leftarrow 1, ..., K_S$
    draw slot dist. $\theta^{ds}_S \sim Dir(\alpha^\star_S)$.
endfor
draw $\phi^s_S \sim Dir(\beta)$ for each slot type $s \leftarrow 1, ..., K_S$.
for each utterance $u \leftarrow 1, ..., |U|$ **do**
  Sample a domain $D_u \sim$Multi(θ^d_D) and,
  and act topic $A_{ud} \sim$Multi(θ^{da}_A).
  for words w_{uj}, $j \leftarrow 1, ..., N_u$ **do**
    - Draw $S_{ujd} \sim$Multi($\theta^{D_u, S_{u(j-1)d}}_S$)‡.
    - Sample $w_{uj} \sim$Multi($\phi^{S_{ujd}}$).
  end for
end for

† $Dir(\alpha^\star_D)$, $Dir(\alpha^\star_A)$, $Dir(\alpha^\star_S)$ are parameterized based on prior knowledge.
‡ Here HMM assumption over utterance words is used.

In hierarchical topic models (Blei et al., 2003; Mimno et al., 2007), etc., topics are represented as distributions over words, and each document expresses an admixture of these topics, both of which have symmetric Dirichlet (*Dir*) prior distributions. Symmetric Dirichlet distributions are often used, since there is typically no prior knowledge favoring one component over another. In the topic model literature, such constraints are sometimes used to deterministically allocate topic assignments to known labels (Labeled Topic Modeling (Ramage et al., 2009)) or in terms of pre-learnt topics encoded as prior knowledge on topic distributions in documents (Reisinger and Paşca, 2009). Similar to previous work, we define a latent topic per each known semantic component label, e.g., five domain topics for five defined domains. Different from earlier work though, we also inject knowledge that we extract from several resources including entity lists from web search query click logs as well as seed labeled training utterances as prior information. We constrain the generation of the semantic components of our model by encoding prior knowledge in terms of

asymmetric Dirichlet topic priors α=(αm_1,...,αm_K) where each *k*th topic has a prior weight α_k=αm_k, with varying base measure **m**=(m_1,...,m_k) [3].

We update parameter vectors of Dirichlet domain prior $\alpha_D^{u\star}$=$\{(\alpha_D \cdot \psi_D^{u1})$,..., $\alpha_D \cdot \psi_D^{uK_D}\}$, where α_D is the concentration parameter for domain Dirichlet distribution and ψ_D^u=$\{\psi_D^{ud}\}_{d=1}^{K_D}$ is the base measure which we obtain from various resources. Because base measure updates are dependent on prior knowledge of corpus words, each utterance u gets a different base measure. Similarly, we update the parameter vector of the Dirichlet dialog act and slot priors $\alpha_A^{u\star}$=$\{(\alpha_A \cdot \psi_A^{u1})$,...,$(\alpha_A \cdot \psi_A^{uK_A})\}$ and $\alpha_S^{u\star}$=$\{(\alpha_S \cdot \psi_S^{u1})$,...,$(\alpha_S \cdot \psi_S^{uK_S})\}$ using base measures ψ_A^u=$\{\psi_A^{ua}\}_{a=1}^{K_A}$ and $\psi_S u$=$\{\psi_S^{us}\}_{s=1}^{K_S}$ respectively.

Before describing base measure update for domain, act and slot Dirichlet priors, we explain the constraining prior knowledge parameters below:

⋆ **Entity List Base Measure**(ψ_E^j): Entity features are indicative of domain and slots and MCM utilizes these features while sampling topics. For instance, entities hotel-name "*Hilton*" and location *"New York"* are discriminative features in classifying *"find nice cheap double room in New York Hilton"* into correct domain (*hotel*) and slot (*hotel-name*) clusters. We represent entity lists corresponding to known domains as multinomial distributions ψ_E^j, where each $\psi_E^{d|j}$ is the probability of entity-word w_j used in the domain d. Some entities may belong to more than one domain, e.g., *"hotel California"* can either be a movie, or song or hotel name.

⋆ **Web n-Gram Context Base Measure** ($\psi_{\mathbf{G}}^j$): As explained in §3, we use the web n-grams as additional information for calculating the base measures of the Dirichlet topic distributions. Normalized word distributions $\psi_{\mathbf{G}}^j$ over domains were used as weights for domain and dialog act base measure.

⋆ **Corpus n-Gram Base Measure** ($\psi_{\mathbf{C}}^j$): Similar to other measures, MCM also encodes n-gram constraints as word-frequency features extracted from labeled utterances. Concretely, we calculate the frequency of vocabulary items given domain-act label pairs from the training labeled utterances and convert there into probability measures over domain-acts. We encode conditional probabilities $\{\psi_{\mathbf{C}}^{ad|j}\}$∈$\psi_{\mathbf{C}}^j$ as multinomial distributions of words over domain-act pairs, e.g., $\psi_{\mathbf{C}}^{ad|j}$= *P*(*d="restaurant", a="make-reservation"|"table"*).

Base measure update: The α-base measures are used to shape Dirichlet priors $\alpha_D^{u\star}$, $\alpha_A^{u\star}$ and $\alpha_S^{u\star}$. We update the base measures of each sampled domain $D_u = d$ given each vocabulary w_j as:

$$\psi_D^{dj} = \begin{cases} \psi_{\mathbf{E}}^{d|j}, & \psi_{\mathbf{E}}^{d|j} > 0 \\ \psi_{\mathbf{G}}^{d|j}, & otherwise \end{cases} \quad (1)$$

In (1) we assume that entities (E) are more indicative of the domain compared to other n-grams (G) and should be more dominant in sampling decision for domain topics. Given an utterance u, we calculate its base measure ψ_D^{ud} =$(\sum_j^{N_u} \psi_D^{dj})/N_u$.

Once the domain is sampled, we update the prior weight of dialog acts $A_{ud} = a$:

$$\psi_A^{aj} = \psi_{\mathbf{C}}^{ad|j} * \psi_{\mathbf{G}}^{d|j} \quad (2)$$

and slot components $S_{ujd} = s$:

$$\psi_S^{sj} = \psi_{\mathbf{E}}^{d|j} \quad (3)$$

Then we update their base measures for a given u as: ψ_A^{ua} =$(\sum_j^{N_u} \psi_A^{aj})/N_u$ and ψ_S^{us} =$(\sum_j^{N_u} \psi_S^{sj})/N_u$.

4.1 Inference and Learning

The goal of inference is to predict the domain, user's act and slot distributions over each segment given an utterance. The MCM has the following set of parameters: domain-topic distributions θ_D^d for each u, the act-topic distributions θ_A^{da} for each domain topic d of u, local slot-topic distributions for each domain θ^S, and ϕ_S^s for slot-word distributions. Previous work (Asuncion et al., 2009; Wallach et al., 2009) shows that the choice of inference method has negligible effect on the probability of testing documents or inferred topics. Thus, we use Markov Chain Monte Carlo (MCMC) method,specifically Gibbs sampling, to model the posterior distribution $P_{\texttt{MCM}}(D_u, A_{ud}, S_{ujd}|\alpha_D^{u\star}, \alpha_A^{u\star}, \alpha_S^{u\star}, \beta)$ by obtaining samples (D_u, A_{ud}, S_{ujd}) drawn from this distribution. For each utterance u, we sample a domain D_u and act A_{ud} and hyper-parameters α_D and α_A and their base measures ψ_D^{ud}, ψ_A^{ua} (from Eq. 1,2):

$$\theta_D^d = \frac{N_u^d + \alpha_D \psi_D^{ud}}{N_u + \alpha_D^{u\star}}; \quad \theta_A^{da} = \frac{N_{a|ud} + \alpha^A \psi_D^{ud}}{N_{ud} + \alpha_A^{u\star}} \quad (4)$$

The N_u^d is the number of occurrences of domain topic d in utterance u, $N_{a|ud}$ is the number of occurrences of act a given d in u. During sampling of a

[3] See (Wallach, 2008) Chapter 3 for analysis of hyper-priors on topic models.

slot state S_{ujd}, we assume that utterance is generated by the HMM model associated with the assigned domain. For each segment w_{uj} in u, we sample a slot state S_{ujd} given the remaining slots and hyper-parameters α_S, β and base measure ψ_S^{us} (Eq. 3) by:

$$p(S_{ujd} = s|\mathbf{w}, \mathbf{D_u}, \mathbf{S}_{-(\mathbf{ujd})}\alpha_{\mathbf{S}}^{\mathbf{u\star}}, \beta) \propto$$
$$\frac{N_{ujd}^k + \beta}{N_{(.)}^k + V\beta} * (N_s^{D_u, S_{u(j-1)d}} + \alpha_S \psi_S^{us}) *$$

$$\frac{N_{S_{u(j+1)d}}^{D_u,s} + \mathbb{I}(S_{uj-1}, s) + \mathbb{I}(S_{uj+1}, s) + \alpha_S \psi_S^{us}}{N_{(.)}^{D_u,s} + \mathbb{I}(S_{uj-1}, s) + K_D \alpha_S^{u\star}} \quad (5)$$

The N_{ujd}^k is the number of times segment w_{uj} is generated from slot state s in all utterances assigned to domain topic d, $N_{s_2}^{D_u,s_1}$ is the number of transitions from slot state s_1 to s_2, where $s_1 \in \{S_{u(j-1)d}, S_{u(j+1)d}\}$, $\mathbb{I}(s_1, s_2)$=1 if slot s_1=s_2.

4.2 Semantic Structure Extraction with MCM

During Gibbs sampling, we keep track of the frequency of draws of domain, dialog act and slot indicating n-grams w_j, in M_D, M_A and M_S matrices, respectively. These n-grams are context bearing words (examples are shown in Fig.1.). For given u the predicted domain d_u^* is determined by:

$$d_u^* = \arg\max_d \tilde{P}(d|u) = \arg\max_d [\theta_D^d * \prod_{j=1}^{N_u} \frac{M_D^{jd}}{M_D}]$$

and predicted dialog act by $\arg\max_a \tilde{P}(a|ud^*)$:

$$a_u^* = \arg\max_a [\theta_A^{d^*a} * \prod_{j=1}^{N_u} \frac{M_A^{ja}}{M_A}] \quad (6)$$

For each segment w_{uj} in u, its predicted slot are determined by $\arg\max_s P(s_j|w_{uj}, d^*, s_{j-1})$:

$$s_{uj}^* = \arg\max_s [p(S_{ujd^*} = s|.) * \prod_{j=1}^{N_u} \frac{Z_S^{js}}{Z_S}] \quad (7)$$

5 Experiments

We performed several experiments to evaluate our proposed approach. Before presenting our results, we describe our datasets as well as two baselines.

5.1 Datasets, Labels and Tags

Our dataset contains utterances obtained from dialogs between human users and our personal assistant system. We use the transcribed text forms of the utterances obtained from (acoustic modeling engine) to train our models [4]. Thus, our dataset contains 18084 NL utterances, 5034 of which are used for measuring the performance of our models. The dataset consists of five domain classes, i.e, *movie, restaurant, hotel, event, other*, 42 unique dialog acts and 41 slot tags. Each utterance is labeled with a domain, dialog act and a sequence of slot tags corresponding to segments in utterance (see examples in Table 1). Table 2 shows sample dialog act and slot labels. Annotation agreement, Kappa measure (Cohen, 1960), was around 85%.

Domain	Sample Dialog Acts (DAs) & Slots
movie	DAs: find-movie/director/actor,buy-ticket Slots: name, mpaa-rating (*g-rated*), date, director/actor-name, award(*oscar winning*)...
hotel	DAs: find-hotel, book-hotel, Slots: name, room-type(*double*), amenities, smoking, reward-program(*platinum elite*)...
restaurant	DAs: find-restaurant, make-reservation, Slots: opening-hour, amenities, meal-type,...
event	DAs: find-event/ticket/performers, get-info.. Slots: name, type(*concert*), performer....

Table 2: List of domains, dialog acts and semantic slot tags of utterance segments. Examples for some slots values are presented in parenthesis as *italicized*.

We pulled a month of web query logs and extracted over 2 million search queries from the movie, hotel, event, and restaurant domains. We also used generic web queries to compile a set of *'other'* domain queries. Our vocabulary consists of n-grams and segments (phrases) in utterances that are extracted using web n-grams and entity lists of §3. We extract distributions of n-grams and entities to inject as prior weights for entity list base ($\psi_{\mathbf{E}}^j$) and web n-gram context base measures ($\psi_{\mathbf{G}}^j$) (see §4).

5.2 Baselines and Experiment Setup

We evaluated two baselines and two variants of our joint SLU approach as follows:

⋆ **Sequence-SLU**: A traditional approach to SLU extracts domain, dialog act and slots as semantic components of utterances using three sequential models. Typically, domain and dialog act detection models are taken as query classification, where a given NL query is assigned domain and act labels. Among supervised query classification meth-

[4] We submitted sample utterances used in our models as additional resource. Due to licensing issues, we will reveal the full train/test utterances upon acceptance of our paper.

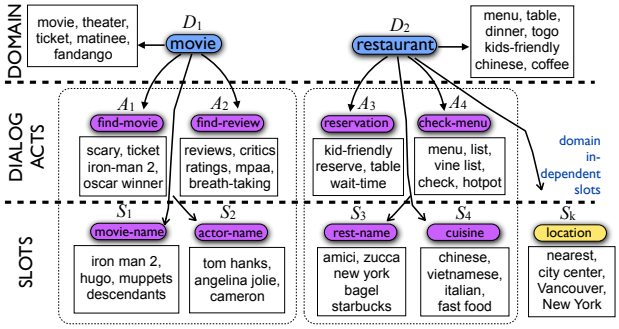


Figure 2: Sample topics discovered by Multi-Layer Context Model (MCM). Given samples of utterances, MCM is able to infer a meaningful set of dialog act (A) and slots (S), falling into broad categories of domain classes (D).

ods, we used the Adaboost, utterance classification method that starts from a set of weak classifiers and builds a strong classifier by boosting the weak classifiers. Slot discovery is taken as a sequence labeling task in which *segments* in utterances are labeled (Li, 2010). For segment labeling we use Semi-Markov Conditional Random Fields (Semi-CRF) (Sarawagi and Cohen, 2004) method as a benchmark in evaluating semantic tagging performance.

⋆ **Tri-CRF**: We used Triangular Chain CRF (Jeong and Lee, 2008) as our supervised joint model baseline. It is a state-of-the art method that learns the sequence labels and utterance class (domain or dialog act) as meta-sequence in a joint framework. It encodes the inter-dependence between the slot sequence $\mathbf{s}$ and meta-sequence label ($\mathbf{d}$ or $\mathbf{a}$) using a triangular chain (dual-layer) structure.

⋆ **Base-MCM**: Our first version injects an informative prior for domain, dialog act and slot topic distributions using information extracted from only labeled training utterances and inject as prior constraints (corpus n-gram base measure $\psi^j_{\mathbf{C}}$) during topic assignments.

⋆ **WebPrior-MCM**: Our full model encodes distributions extracted from labeled training data as well as structured web logs as asymmetric Dirichlet priors. We analyze performance gain by the information from web sources ($\psi^j_{\mathbf{G}}$ and $\psi^j_{\mathbf{E}}$) when injected into our approach compared to Base-MCM.

We inject dictionary constraints as features to train supervised discriminative methods, i.e., boosting and Semi-CRF in Sequence-SLU, and Tri-CRF models. For semantic tagging, dictionary constraints apply to the features between individual segments and their labels, and for utterance classification (to predict domain and dialog acts) they apply to the features between utterance and its label. Given a list of dictionaries, these constraints specify which label is more likely. For discriminative methods, we use several named entities, e.g., Movie-Name, Restaurant-Name, Hotel-Name, etc., non-named entities, e.g., Genre, Cuisine, etc., and domain independent dictionaries, e.g., Time, Location, etc.

We train domain and dialog act classifiers via Icsiboost (Favre et al., 2007) with 10K iterations using lexical features (up to 3-n-grams) and constraining dictionary features (all dictionaries). For feature templates of sequence learners, i.e., Semi-CRF and Tri-CRF, we use current word, bi-gram and dictionary features. For Base-MCM and WebPrior-MCM, we run Gibbs sampler for 2000 iterations with the first 500 samples as burn-in.

5.3 Evaluations and Discussions

We evaluate the performance of our joint model on two experiments using two metrics. For domain and dialog act detection performance we present results in accuracy, and for slot detection we use the F1 pairwise measure.

Experiment 1. Encoding Prior Knowledge: A common evaluation method in SLU tasks is to measure the performance of each individual semantic model, i.e., domain, dialog act and semantic tagging (slot filling). Here, we not only want to demonstrate the performance of each component of MCM but also their performance under limited amount of labeled data. We randomly select subsets of labeled training data $U^i_L \subset U_L$ with different samples sizes, $n^i_L = \{\gamma * n_L\}$, where n_L represents the sample size of U_L and γ=$\{10\%,25\%,..\}$ is the subset percentage. At each random selection, the rest of the utterances are used as unlabeled data to boost the performance of MCM. The supervised baselines do not leverage the unlabeled utterances.

The results reported in Figure 3 reveal both the strengths and some shortcomings of our approach. When the number of labeled data is small ($n^i_L \leq 25\% * n_L$), our WebPrior-MCM has a better performance on domain and act predictions compared to the two baselines. Compared to Sequence-SLU, we observe 4.5% and 3% performance improvement on the domain and dialog act

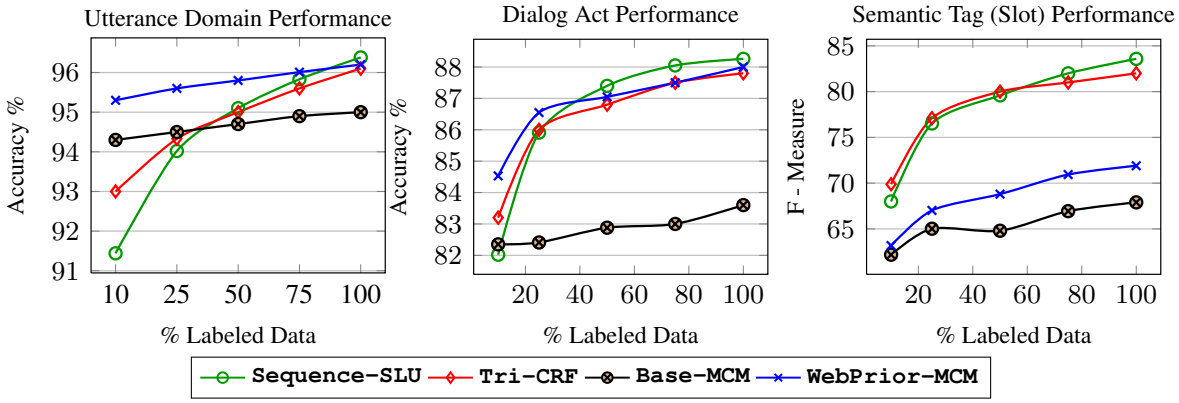


Figure 3: Semantic component extraction performance measures for various baselines as well as our approach with different priors.

models, whereas our gain is 2.6% and 1.7% over Tri-CRF models. As the percentage of labeled utterances in training data increase, Tri-CRF performance increases, however WebPrior-MCM is still comparable with Sequence-SLU. This is because we utilize domain priors obtained from the web sources as supervision during generative process as well as unlabeled utterances that enable handling language variability. Adding labeled data improves the performance of all models however supervised models benefit more compared to MCM models.

Although WebPrior-MCM's domain and dialog act performances are comparable (if not better than) the other baselines, it falls short on the semantic tagging model. This is partially due to the HMM assumption compared to the supervised conditional model's used in the other baselines, i.e., Semi-CRF in Sequence-SLU and Tri-CRF). Our work can be extended by replacing HMM assumption with CRF based sequence learner to enhance the capability of the sequence tagging component of MCM.

Experiment 2. Less is More? Being Bayesian, our model can incorporate unlabeled data at training time. Here, we evaluate the performance gain on domain, act and slot predictions as more unlabeled data is introduced at learning time. We use only 10% of the utterances as labeled data in this experiment and incrementally add unlabeled data (90% of labeled data are treated as unlabeled).

The results are shown in Table 3. $n\%$ (n=10,25,..) unlabeled data indicates that the WebPrior-MCM is trained using $n\%$ of unlabeled utterances along with training utterances. Adding unlabeled data has a positive impact on the performance of all three semantic components when WebPrior-MCM is used. The results show that our joint modeling approach has an advantage over the other joint models (i.e., Tri-CRF) in that it can leverage unlabeled NL utterances. Our approach might be usefully extended into the area of understanding search queries, where an abundance of unlabeled queries is observed.

Table 3: Performance evaluation results of WebPrior-MCM using different sizes of unlabeled utterances at learning time.

Unlabeled %	Domain Accuracy	Dialog Act Accuracy	Slot F-Measure
10%	94.69	84.17	52.61
25%	94.89	84.29	54.22
50%	95.08	84.39	56.58
75%	95.19	84.44	**57.45**
100%	**95.28**	84.52	**58.18**

6 Conclusions

In this work, we introduced a joint approach to spoken language understanding that integrates two properties (*i*) identifying user actions in multiple domains in relation to semantic units, (*ii*) utilizing large amounts of unlabeled web search queries that suggest the user's hidden intentions. We proposed a semi-supervised generative joint learning approach tailored for injecting prior knowledge to enhance the semantic component extraction from utterances as a unifying framework. Experimental results using the new Bayesian model indicate that we can effectively learn and discover meta-aspects in natural language utterances, outperforming the supervised baselines, especially when there are fewer labeled and more unlabeled utterances.

References

A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. 2009. On smoothing and inference for topic models. *UAI*.

S. Bangalore. 2006. Introduction to special issue of spoken language understanding in conversational systems. In *Speech Conversation*, volume 48, pages 233–238.

L. Begeja, B. Renger, Z. Liu D. Gibbon, and B. Shahraray. 2004. Interactive machine learning techniques for improving slu models. In *Proceedings of the HLT-NAACL 2004 Workshop on Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Information for Speech Processing*.

Pauline M. Berry, Melinda Gervasio, Bart Peintner, and Neil Yorke-Smith. 2011. Ptime: Personalized assistance for calendaring. In *ACM Transactions on Intelligent Systems and Technology*, volume 2, pages 1–40.

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

J. Cohen. 1960. A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement*, volume 20, pages 37–46.

H. Daumé-III and D. Marcu. 2006. Bayesian query focused summarization.

M. Dinarelli, A. Moschitti, and G. Riccardi. 2009. Reranking models for spoken language understanding. *Proc. European Chapter of the Annual Meeting of the Association of Computational Linguistics (EACL)*.

B. Favre, D. Hakkani-Tür, and Sebastien Cuendet. 2007. Icsiboost. `http://code.google.come/p/icsiboost`.

S. Harabagiu and A. Hickl. 2006. Methods for using textual entailment for question answering. pages 905–912.

E. Hovy, C.Y. Lin, and L. Zhou. 2005. A be-based multi-document summarizer with query interpretation. *Proc. DUC*.

M. Jeong and G. G. Lee. 2008. Triangular-chain conditional random fields. *EEE Transactions on Audio, Speech and Language Processing (IEEE-TASLP)*.

X. Li. 2010. Understanding semantic structure of noun phrase queries. *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*.

P. Liang, M. I. Jordan, and D. Klein. 2011. Learning dependency based compositional semantics.

A. Margolis, K. Livescu, and M. Osterdorf. 2010. Domain adaptation with unlabeled data for dialog act tagging. In *Proc. Workshop on Domain Adaptation for Natural Language Processing at the the Annual Meeting of the Association of Computational Linguistics (ACL)*.

D. Mimno, W. Li, and A. McCallum. 2007. Mixtures of hierarchical topics with pachinko allocation. *Proc. ICML*.

A. Popescu, P. Pantel, and G. Mishne. 2010. Semantic lexicon adaptation for use in query interpretation. *19th World Wide Web Conference (WWW-10)*.

D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. *Proc. EMNLP*.

J. Reisinger and M. Paşca. 2009. Latent variable models of concept-attribute attachement. *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*.

J. Reisinger and M. Pasca. 2011. Fine-grained class label markup of search queries. In *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*.

M. Sammons, V. Vydiswaran, and D. Roth. 2010. Ask not what textual entailment can do for you... In *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*, Uppsala, Sweden, 7.

S. Sarawagi and W. W. Cohen. 2004. Semimarkov conditional random fields for information extraction. *Proc. NIPS*.

C. Sauper, A. Haghighi, and R. Barzilay. 2011. Content models with attitude. In *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*.

G. Tur and R. De Mori. 2011. Spoken language understanding: Systems for extracting semantic information from speech. *Wiley*.

H. Wallach, D. Mimno, and A. McCallum. 2009. Rethinking lda: Why priors matter. *NIPS*.

H. Wallach. 2008. Structured topic models for language. *Ph.D. Thesis, University of Cambridge*.

Y.Y. Wang, R. Hoffman, X. Li, and J. Syzmanski. 2009. Semi-supervised learning of semantic classes for query understanding from the web and for the web. In *The 18th ACM Conference on Information and Knowledge Management*.

Y-Y. Wang. 2010. Strategies for statistical spoken language understanding with small amount of data - an emprical study. *Proc. Interspeech 2010*.