Predicting Tasks in Goal-Oriented Spoken Dialog Systems using Semantic Knowledge Bases

Aasish Pappu and Alexander I. Rudnicky
Language Technologies Institute
Carnegie Mellon University
{aasish, air}@cs.cmu.edu

Abstract

Goal-oriented dialog agents are expected to recognize user-intentions from an utterance and execute appropriate tasks. Typically, such systems use a semantic parser to solve this problem. However, semantic parsers could fail if user utterances contain out-of-grammar words/phrases or if the semantics of uttered phrases did not match the parser's expectations. In this work, we have explored a more robust method of task prediction. We define task prediction as a classification problem, rather than "parsing" and use semantic contexts to improve classification accuracy. Our classifier uses semantic smoothing kernels that can encode information from knowledge bases such as Wordnet, NELL and Freebase.com. Our experiments on two spoken language corpora show that augmenting semantic information from these knowledge bases gives about 30% absolute improvement in task prediction over a parser-based method. Our approach thus helps make a dialog agent more robust to user input and helps reduce number of turns required to detected intended tasks.

1 Introduction

Spoken dialog agents are designed with particular tasks in mind. These agents could provide information or make reservations, or other such tasks. Many dialog agents often can perform multiple tasks: think of a customer service kiosk system at a bank. The system has to decide which task it has to perform by talking to its user. This problem of identifying what to do based on what a user has said is called task prediction.

Task prediction is typically framed as a parsing problem: A grammar is written to semantically parse the input utterance from users, and these semantic labels in combination are used to decide what the intended task is. However, this method is less robust to errors in user-input. A dialog system consists of a pipeline of cascaded modules, such as speech recognition, parsing, dialog management. Any errors made by these modules propogate and accumulate through the pipeline. Bohus and Rudnicky (2005) have shown that this cascade of errors, coupled with users employing out-of-grammar phrases results in many "non-understanding" and "misunderstanding" errors.

There have been other approaches to perform dialog task prediction. Gorin et al. (1997) has proposed a salience-phrase detection technique that maps phrases to their corresponding tasks. Chu-Carroll and Carpenter (1999) casted the task detection as an information retrieval — detect tasks by measuring the distance between the query vector and representative text for each task. Bui (2003) and Blaylock and Allen (2006) have cast it as a hierarchical sequence labeling problem using Hidden Markov Models (HMM). More recently, (Bangalore and Stent, 2009) built an incremental parser that gradually determines the task based on the incoming dialog utterances. (Chen and Mooney, 2010) have developed a route instructions frame parser to determine the task in the context of a mobile dialog robot. These approaches mainly use local features such as dialog context, speech features and grammar-based-semantic features to determine the task. However grammar-based-semantic features would be insufficient if an utterance uses semantically similar phrases that are not in the system's domain or semantics. If the system could explore semantic information beyond the scope of its local knowledge and use external knowledge sources then they will help improve the task prediction.

(Cristianini et al., 2002) (Wang and Domeniconi, 2008) (Moschitti, 2009) found that open-

Proceedings of the SIGDIAL 2013 Conference, pages 242–250,
Metz, France, 22-24 August 2013.

domain semantic knowledge resources are useful for text classification problems. Their success in limited data scenario is an attractive prospect, since most dialog agents operate in scarce training data scenarios. (Bloehdorn et al., 2006) has proposed a semantic smoothing kernel based approach for text classification. The intuition behind their approach is that terms (particularly content words) of two similar sentences or documents share superconcepts (e.g., hypernyms) in a knowledge base. Semantic Similarity between two terms can be computed using different metrics (Pedersen et al., 2004) based on resources like WordNet.

Open domain resources such as world-wide-web, had been used in the context of speech recognition. (Misu and Kawahara, 2006) and (Creutz et al., 2009) used web-texts to improve the language models for speech recognition in a target domain. They have used a dialog corpus in order to query relevant web-texts to build the target domain models. Although (Araki, 2012) did not conduct empirical experiments, yet they have presented an interesting architecture that exploits an open-domain resource like Freebase.com to build spoken dialog systems.

In this work, we have framed the task prediction problem as a classification problem. We use the user's utterances to extract lexical semantic features and classify it into being one of the many tasks the system was designed to perform. We harness the power of semantic knowledge bases by bootstraping an utterance with semantic concepts related to the tokens in the utterance. The semantic distance/similarity between concepts in the knowledge base is incorporated into the model using a kernel. We show that our approach improves the task prediction accuracy over a grammar-based approach on two spoken corpora (1) Navagati (Pappu and Rudnicky, 2012): a corpus of spoken route instructions, and (2) Roomline (Bohus, 2003): a corpus of spoken dialog sessions in room-reservation domain.

This paper is organized as following: Section 2 describes the problem of dialog task prediction and the standard grammar based approach to predict the dialog task. Then in Section 3, we describe the open-domain knowledge resources that were used in our approach and their advantages/disadvantages. We will discuss our semantic kernel based approach in the Section 4. We report our experiment results on task prediction in Section 5. In Section 6, we will analyze the errors that occur in our approach, followed by concluding remarks and possible directions to this work.

2 Parser based Dialog Task Prediction

In a dialog system, there are two functions of a semantic grammar — encode linguistic constructs used during the interactions and represent the domain knowledge in-terms of concepts and their instances. Table 1 illustrates the tasks and the concepts used in a navigation domain grammar. The linguistic constructions help the parser to segment an utterance into meaningful chunks. The domain knowledge helps in labeling the tokens/phrases with concepts. The parser uses the labeled tokens and the chunked form of the utterance, to classify the utterance into one of the tasks.

Table 1: Tasks and Concepts in Grammar

Tasks	**Examples**
Imperative	GoToPlace, Turn, etc
Advisory Instructions	You_Will_See_Location
Grounding Instructions	You_are_at_Location
Concepts	**Examples**
Locations	buildings, other landmarks
Adjectives-of-Locations	large, open, black, small etc.
Pathways	hallway, corridor, bridge, etc.
LiftingDevice	elevator, staircase, etc.
Spatial Relations	behind, above, on left, etc.
Numbers	turn-angles, distance, etc.
Ordinals	first, second, etc. floor numbers

The dialog agent uses the root node of a parser output as the task. Figure 1 illustrates a semantic parser output for a fictitious utterance in the navigation domain. The dialog manager would consider the utterance as an "Imperative" for this example.

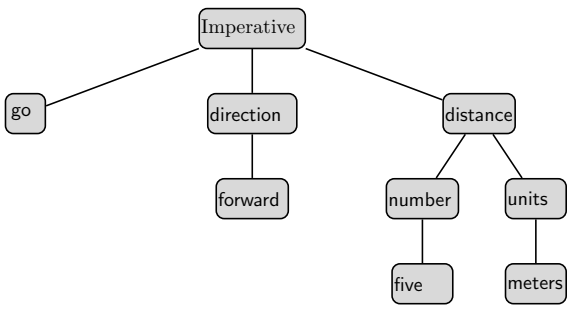

Figure 1: Illustration of Semantic Parse Tree used in a Dialog System

2.1 Grammar: A Knowledge Resource

Grammar is a very useful resource for a dialog system because it could potentially represent an expert's view of the domain. Since knowledge engineering requires time and effort, very few dialog systems can afford to have grammars that are expert-crafted and robust to various artefacts of spoken language. This becomes a major challenge for real world dialog systems. If the system's grammar or the domain knowledge does not conform to its users and their utterances, the parser will fail to produce a correct parse, if the parse is incorrect and/or the concept labeling is incorrect. Lack of comprehensive semantic knowledge is the cause of this problem. An open-domain knowledge base like Wordnet (Miller, 1995), Freebase (Bollacker et al., 2008) or NELL (Carlson et al., 2010) contains comprehensive information about concepts and their relationships present in the world. If used appropriately, open-domain knowledge resources can help compensate for incomplete semantic knowledge of the system.

3 Open-Domain Semantic Knowledge Bases

Like grammars, open-domain knowledge resources contain concepts, instances and relations. The purpose of these resources is to organize common sense and factoid information known to the mankind in a machine-understandable form. These resources, if filtered appropriately, contain valuable domain-specific information for a dialog agent. To this end, we propose to use three knowledge resources along with the domain grammar for the task prediction. A brief overview of each of the knowledge resources is given below:

3.1 Wordnet: Expert Knowledge Base

Wordnet (Miller, 1995) is an online lexical database of words and their semantics curated by language experts. It organizes the words and their morphological variants in a hierarchical fashion. Every word has at least one synset i.e., sense and a synset has definite meaning and a gloss to illustrate the usage. Synsets are connected through relationships such as hypernyms, hyponyms, meronyms, antonyms etc. Each synset can be considered as an instance and their parent synsets as concepts. Although Wordnet contains several (120,000) word forms, some of our domain-specific word forms (e.g., locations in a navigation domain) will not be present. Therefore, we would like to use other open-domain knowledge bases to augment the agent's knowledge.

3.2 Freebase: Community Knowledge Base

Freebase.com (Bollacker et al., 2008) is a collaboratively evolving knowledge base with the effort of volunteers. It organizes the facts based on types/concepts along with several predicates/properties and their values for each fact. The types are arranged in a hierarchy and the hierarchy is rooted at "domain". Freebase facts are constantly updated by the volunteers. Therefore, it is a good resource to help bootstrap the domain knowledge of a dialog agent.

3.3 NELL: Automated Knowledge Base

Never-Ending Language Learner(NELL) (Carlson et al., 2010) is a program that learns and organizes the facts from the web in an unsupervised fashion. NELL is on the other end of the knowledge base spectrum which is not curated either by experts or by volunteers. NELL uses a two-step approach to learn new facts: (1) extract information from the text using pattern-based, semi-structured relation extractors (2) improve the learning for next iteration based on the evidence from previous iteration. Every belief/fact in its knowledge base has concepts, source urls, extraction patterns, predicate, the surface forms of the facts and a confidence score for the belief. Although the facts could be noisy in comparison to ones in other knowledge bases, NELL continually adds and improves the facts without much human effort.

4 Semantic Kernel based Dialog Task Prediction

We would like to use this apriori knowledge about the world and the domain to help us predict the dialog task. The task prediction problem can be treated as a classification problem. Classification algorithms typically use bag-of-words representation that converts a document or sentence into a vector with terms as components of the vector. This representation produces very good results in scenarios with sufficient training data. However in a limited training data or extreme sparseness scenario such as ours, (Siolas and d'Alché Buc, 2000) has shown that Semantic Smoothing Kernel technique is a promising approach. The major advantage of this approach is that they can incor-

porate apriori knowledge from existing knowledge bases. The semantic dependencies between terms, dependencies between concepts and instances, can be encoded in these kernels. The semantic kernels can be easily plugged into a kernel based classifier help us predict the task from the goal-oriented dialog utterances.

In our experiments, we used an implementation of Semantic Kernel from (Bloehdorn et al., 2006) and plugged it into a Support Vector Machine (SVM) classifier (SVM^{light}) (Joachims, 1999). As a part of experimental setup, we will describe the details of how did we extract the semantic dependencies from each knowledge base and encoded them into the kernel.

5 Experiments

Our goal is to improve the task prediction for a given spoken dialog utterance by providing additional semantic context to the utterance with the help of relevant semantic concepts from the semantic knowledge bases. The baseline approach would use the Phoenix parser's output to determine the intended task for an utterance. From our experiments, we show that our knowledge-driven approach will improve upon the baseline performance on two corpora (1) Navagati Corpus: a navigation directions corpus (2) Roomline Corpus: a room reservation dialog corpus.

5.1 Setup

We have divided each corpus into training and testing datasets. We train our task classification models on the manual transcriptions of the training data and evaluated the models on the ASR output of the testing data. Both Navagati and Roomline corpora came with manually annotated task labels and manual transcriptions for the utterances. We filtered out the non-task utterances such as "yes", "no" and other clarifications from the Roomline corpus. We obtained the ASR output for the Navagati corpus by running the test utterances through PocketSphinx (Huggins-Daines et al., 2006). The Roomline corpus already had the ASR output for the utterances. Table 2 illustrates some of the statistics for each corpus.

Our baseline model for the task detection is the Phoenix (Ward, 1991) parser output, which is the default method used in the Ravenclaw/Olympus dialog systems (Bohus et al., 2007). For the Navagati Corpus we have obtained the parser output using the grammar and method described in (Pappu and Rudnicky, 2012). For the Roomline corpus, we extracted the parser output from the session logs from the the corpus distribution.

Corpus-Stats	**Navagati**	**RoomLine**
Tasks	4	7
Words	503	498
Word-Error-rate	46.3%	25.6%
Task Utts	934	1891[1]
Task Training-Utts	654	1324
Task Testing-Utts	280	567
Tasks		
	N1. Meta N2. Advisory N3. Imperative N4. Grounding	R1. NeedRoom R2. ChooseRoom R3. QueryFeatures R4. ListRooms R5. Identification R6. CancelReservation R7. RejectRooms

Table 2: Corpus Statistics

5.1.1 Semantic Facts to Semantic Kernel

The semantic kernel takes a term proximity matrix as an input, then produces a positive semidefinite matrix which can be used inside the kernel function. In our case, we build a term proximity matrix for the words in the recognition vocabulary. (Bloehdorn et al., 2006) found that using the term-concept pairs in the proximity matrix is more meaningful following the intuition that terms that share more number of concepts are similar as opposed to terms that share fewer concepts. We have used following measures to compute the proximity value P and some of them are specific to respective knowledge bases:

- `gra`: No weighting for term-concept pairs in the Grammar, i.e.,
  $P = 1$, for all concepts c_i of t, $P = 0$ otherwise.
- `fb`: Weighting using normalized Freebase.com relevance score, i.e.,

$$P = \frac{fbscore(t, c_i) - fbscore(t, c_{min})}{fbscore(t, c_{max}) - fbscore(t, c_{min})} \quad (1)$$

- `nell`: Weighting for the NELL term-concept pairs using the probability for a belief i.e.,

$$P = nellprob(t, c_i) \quad (2)$$

, for all concepts c_i of t, $P = 0$ otherwise.

[1] Originally has 10356 utts; filtered out non-task utts.

- `wnpath`: Weighting for the term-concept pairs in the Wordnet based on the shortest path, i.e.,

$$P = wn_{PATH}(t, c_i) \quad (3)$$

for all concepts c_i of t, $P = 0$ otherwise.

- `wnlch`: Weighting for the term-concept pairs in the Wordnet based on the Leacock-Chodorow Similiarity, i.e.,

$$P = wn_{LCH}(t, c_i) \quad (4)$$

for all concepts c_i of t, $P = 0$ otherwise.

- `wnwup`: Weighting for the term-concept pairs in the Wordnet based on the Wu-Palmer Similarity, i.e.,

$$P = wn_{WUP}(t, c_i) \quad (5)$$

for all concepts c_i of t, $P = 0$ otherwise.

- `wnres`: Weighting for the term-concept pairs in the Wordnet based on the Resnik Similarity using Information Content, i.e.,

$$P = wn_{RES}(t, c_i) \quad (6)$$

for all concepts c_i of t, $P = 0$ otherwise.

To create a grammar-based proximity matrix, we extracted the concept-token pairs from the parser output on the reference transcriptions in both corpora. In order to create a wordnet-based proximity matrix, we retrieve the hypernyms for the corresponding from Wordnet using the Wordnet 3.0 database [2]. For the freebase concept-token pairs, we query tokens for a list of types with the help of the MQL query interface[3] to the freebase. To retrieve beliefs from NELL we downloaded a tsv formatted database called every-belief-in-the-KB[4] and then queried for facts using unix `grep` command.

5.2 Results

Our objective is to evalute the effect of augmented semantic features on the task detection. As noted earlier, we divided both corpora into training and testing datasets. We build our models on the manual transcriptions from the training data and evaluate on the ASR hypotheses of the testing data. For the Navagati corpus, we use the same training-testing split that we used in our previous work because the grammar was developed based on the training data. For the Roomline corpus, we randomly sample 30% of the testing data from the entire corpus.

Our first semantic-kernel based model SEM-GRA uses the domain grammar as a "knowledge base". This is a two step process: (1) we extract the concept-token pairs from the parse output of the training data. (2) Then, assign a uniform proximity score (1.0) for all pairs of words that appear under a particular concept otherwise 0.0 (`gra` as mentioned in the previous section). We augment the grammar concepts to the utterances in the datasets, learn SEM-GRA model and classify the test-hypotheses. For all our models we use a fixed $C = 0.07$ value (soft-margin parameter) for the SVM classifiers. This model achieved highest performance at this value during a parameter-sweep. SEM-GRA model outperformed the parser-baseline on both datasets (see Table 3). It clearly takes advantage of the domain knowledge encoded in the form of semantic-relatedness between concepts and token pairs.

What if a dialog system does not have grammar to begin with? We use the same two step process to build semantic-kernel based models using one open-domain knowledge base at a time. We built Wordnet based models (SEM-WNWUP, SEM-WNPATH, SEM-WNLCH, SEM-WNRES) using different proximity measures described in the previous section. From Table 3 SEM-WNRES model, one that uses information content performs the best among all wordnet based models. In order to compute the information content we used the pair-wise mutual information scores available for brown-corpus.dat in the NLTK (Bird et al., 2009) distribution. Other path based scores were also computed using NLTK API for Wordnet.

We observed that our wordnet-based models capture relatedness between most-common nouns (e.g., room numbers) and their concepts but not for some of the less-common ones (e.g., location names). To compensate this imbalance, we use larger knowledge resources freebase.com and NELL. First we searched for the facts in each of these knowledge bases using every token in the vocabulary of both corpora. We pick the top concept for each token based on the score provided by the respective search interfaces. In freebase we have

[2] http://www.princeton.edu/wordnet/download/
[3] https://www.googleapis.com/freebase/v1/search
[4] http://rtw.ml.cmu.edu/rtw/resources

Table 3: F1 (in %) comparison of parse baseline against semantic-kernel models with their corresponding similarity metrics

Corpus	baseline	SEMGRA	SEMWNWUP	SEMWNPATH	SEMWNLCH	SEMWNRES	SEMFBASE	SEMNELL
Navagati	40.1	65.8	67.1	67.7	66.4	**69**	68.7	66.2
Roomline	54.3	79.7	77.3	79.5	79.6	**80.6**	83.3	81.1

about 100 concepts that are relevant to the vocabulary and in the NELL model we have about 250 concepts that are relevant to the vocabulary in each of the corpora. The models based on NELL (SEM-NELL) and Freebase (SEM-FBASE) capture relatedness between less-common nouns and their concepts. We can see that both of these models perform comparable to the domain grammar model SEM-GRA which also captures the relatedness between less-common nouns and their concepts. We believe that both freebase and NELL has a superior performance because of wider-range of concept coverage and non-uniform proximity measures used in the semantic kernel, which gives a better judgement of relatedness than a uniform measure used in the SEM-GRA model.

Since we observed that an individual model is good at capturing a particular aspect of an utterance, we extended the individual semantic models by combining the proximity matrices from each of them and augmenting their semantic concepts to the training and testing datasets. We built four combined models as shown in Table 4 by varying the wordnet's proximity metric to identify which one of them works best in combination with other semantic metrics. The `wnres` metric performs the best both in standalone and combination settings. (Bloehdorn et al., 2006) also found that `wnres` particularly performs well for lower values of the soft-margin parameter in their experiments.

Table 4: F1-Score (in %): Models with semantics combined from different KBs (ALL-KB)

Model	**Navagati**	**Roomline**
GRA+WNWUP+FBASE+NELL	70.8	82.2
GRA+WNPATH+FBASE+NELL	70.1	81.4
GRA+WNLCH+FBASE+NELL	70.8	81.3
GRA+WNRES+FBASE+NELL	**73.4**	**83.7**

6 Discussion

We have built a model that exploits different semantic knowledge bases and predicts the task on both corpora with high accuracy. But how is it affected by factors like misrecognition and context ambiguity?

6.1 Influence of Recognition Errors

When the recognition is bad, it is obvious that the accuracy would go down. We would like to know which of these knowledge resources can augment useful semantics despite misrecognitions. Table 2 shows that WER on the Navagati corpus is about 46% and the Roomline corpus is about 25%. We compared the F1-score of different models on utterances for different ranges of WER as shown in the Figure 2 on the Navagati Corpus. We notice that the model built using all knowledge bases is more robust even at higher WER. We did similar analysis on the Roomline corpus and did not notice any differences across models due to relatively lower WER (25.6%).

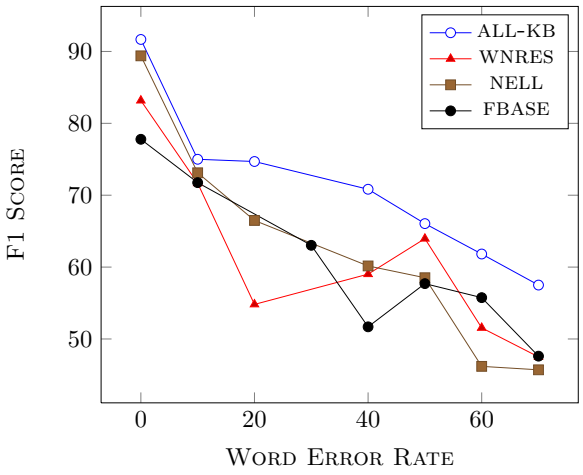

Figure 2: Word Error Rate vs F1-Score for KB-based Models on Navagati Corpus

6.2 Confusion among Tasks

We found that a particular pair of tasks are more confusing than others. Here we present an analysis of such confusion pairs for both corpora for different classification models. Table 5 and Table 6 show the pairs of tasks that are most confused in the experiments. The ALL-KB model (a combination of all knowledge bases) has least number of

Table 5: Most confusable pairs of tasks in Navagati Corpus for KB based classification models (See Table 2 for task labels)

KBType	ALL-KB		SEM-WNRES		SEM-NELL		SEM-FBASE		
ActualTask	N2	N4	N2	N4	N2	N4	N1	N2	N4
Predicted	N3	N1	N3	N3	N3	N3	N3	N3	N3
ConfusionPerTask	10.5%	27.7%	26.3%	33.3%	26.3%	38.8%	22.2%	28.9%	44.4%

Table 6: Most confusable pairs of tasks in Roomline Corpus for KB based classification models (See Table 2 for task labels)

KBType	ALL-KB	SEM-WNRES		SEM-NELL		SEM-FBASE			
ActualTask	R4	R4	R6	R4	R6	R3	R4	R5	R6
Predicted	R3	R5	R5	R1	R1	R1	R3	R1	R1
ConfusionPerTask	36.6%	48.7%	44.4%	25.6%	44.5%	32.5%	23%	53.4%	55.5%

confusion pairs among all the models. This is due to more relevant concepts are augmented to an utterance compared to fewer relevant concepts that augmented while using individual models.

We inspected the confused tasks by examining the feature vectors of misclassified examples. While using the ALL-KB model 10% of the utterances from N2 (Advisory) were confused for N3 (Imperative) because of phrases like "your left", "your right". These phrases were often associated with N3 utterances. To recovery from such ambiguities, the agent could ask a clarification question e.g., "are we talking about going there or find it on the way?" to resolve the differences between these tasks. The system could not only get clarification but also bootstrap the original utterance of the user with the clarification to gather additional context to retrain the task detection models. The individual models were also confused about N2 and N3 tasks, where we could use similar clarification strategies to improve the task prediction. 27% of the N4 (grounding about current robot's position) utterances were confused for N1 (meta comments about the robot's rounavigation route) because these utterances shared more number of freebase concepts with the N1 model. The system could resolve such utterances by asking a clarification question "are we talking about the current position?". Individual models i.e., SEM-WNRES, SEM-FBASE and SEM-NELL suffered mostly from the lack of concepts capturing semantics related to all types of entities (e.g., most common nouns, less common entities etc.,) found in an utterance.

We examined the confusion pairs in the Roomline corpus and observed that R4 (ListRooms) and R3 (Queries) tasks were most confused in the ALL-KB model. On closer inspection, we found that R4 utterances are about listing the rooms that are retrieved by the system. Whereas, R3 utterances are about asking whether a room has a facility (e.g., projector availability). In the ambiguous utterances, often the R4 utterances were about filtering the list of rooms by a facility type.

7 Conclusion

We proposed framing the dialog task prediction problem as a classification problem. We used an SVM classifier with semantic smoothing kernels that incorporate information from external knowledge bases such as Wordnet, NELL, Freebase. Our method shows good improvements over a parser-based baseline. Our analysis also shows that our proposed method makes task prediction be more robust to moderate recognition errors.

We presented an analysis on task ambiguity and found that these models can confuse one task for another. We believe that this analysis highlights the need for dialog based clarification strategies that cannot only help the system for that instance but also help the system improve its task prediction accuracy in future dialog sessions.

8 Future Work

This work stands as a platform to make a spoken dialog system learn relevant semantic information from external knowledge sources. We would like to extend this paradigm to let the system acquire more information through dialog with a user. The system could elicit new references to a known semantic concept. For example, a navigation agent knows a task called "GoToRestaurant" but the user-utterance had the word "diner" and it was

not seen in the context of "restaurant". The agent somewhat predicts this utterance is related to "GoToRestaurant" using the approach described in this paper. It could ask the user an elicitation question: "You used diner in the context of a restaurant, is diner really a restaurant?". The answer to this question will help the system gradually understand what parts of an open-domain knowledge base can be added into its own domain knowledge base. We believe that the holistic approach of learning from automated processes and learning through dialog, will help the dialog systems get better interaction by interaction.

References

Masahiro Araki. 2012. Rapid development process of spoken dialogue systems using collaboratively constructed semantic resources. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 70–73, Seoul, South Korea, July. Association for Computational Linguistics.

Srinivas Bangalore and Amanda J Stent. 2009. Incremental parsing models for dialog task structure. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media.

Nate Blaylock and James Allen. 2006. Hierarchical instantiated goal recognition. In *Proceedings of the AAAI Workshop on Modeling Others from Observations*.

Stephan Bloehdorn, Roberto Basili, Marco Cammisa, and Alessandro Moschitti. 2006. Semantic kernels for text classification based on topological measures of feature similarity. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 808–812. IEEE.

Dan Bohus and Alexander I Rudnicky. 2005. Sorry, I didn't catch that!-an investigation of non-understanding errors and recovery strategies. In *6th SIGdial Workshop on Discourse and Dialogue*.

Dan Bohus, Antoine Raux, Thomas K Harris, Maxine Eskenazi, and Alexander I Rudnicky. 2007. Olympus: an open-source framework for conversational spoken language interface research. In *Proceedings of the workshop on bridging the gap Academic and industrial research in dialog technologies*, number April, pages 32–39. Association for Computational Linguistics.

Dan Bohus. 2003. Roomline. `http://www.cs.cmu.edu/~dbohus/RoomLine`.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. *SIGMOD 08 Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1249.

Hung H Bui. 2003. A general model for online probabilistic plan recognition. In *International Joint Conference on Artificial Intelligence*, volume 18, pages 1309–1318. Citeseer.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr., and Tom M Mitchell. 2010. Toward an Architecture for Never-Ending Language Learning. *Artificial Intelligence*, 2(4):1306–1313.

D.L. Chen and R.J. Mooney. 2010. Learning to interpret natural language navigation instructions from observations. *Journal of Artificial Intelligence Research*, 37:397–435.

Jennifer Chu-Carroll and Bob Carpenter. 1999. Vector-based natural language call routing. *Computational linguistics*, 25(3):361–388.

Mathias Creutz, Sami Virpioja, and Anna Kovaleva. 2009. Web augmentation of language models for continuous speech recognition of sms text messages. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 157–165. Association for Computational Linguistics.

Nello Cristianini, John Shawe-Taylor, and Huma Lodhi. 2002. Latent semantic kernels. *Journal of Intelligent Information Systems*, 18(2):127–152.

Allen L Gorin, Giuseppe Riccardi, and Jeremy H Wright. 1997. How may i help you? *Speech communication*, 23(1-2):113–127.

D. Huggins-Daines, M. Kumar, A. Chan, A.W. Black, M. Ravishankar, and A.I. Rudnicky. 2006. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *ICASSP*, volume 1. IEEE.

Thorsten Joachims. 1999. Making large-scale support vector machine learning practical. In *Advances in kernel methods*, pages 169–184. MIT Press.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Teruhisa Misu and Tatsuya Kawahara. 2006. A bootstrapping approach for developing language model of new spoken dialogue systems by selecting web texts. In *Proc. Interspeech*, pages 9–12.

Alessandro Moschitti. 2009. Syntactic and semantic kernels for short text pair categorization. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 576–584.

Aasish Pappu and Alexander I Rudnicky. 2012. The Structure and Generality of Spoken Route Instructions. *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 99–107.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics.

Georges Siolas and Florence d'Alché Buc. 2000. Support vector machines based on a semantic kernel for text categorization. In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference*, volume 5, pages 205–209. IEEE.

Pu Wang and Carlotta Domeniconi. 2008. Building semantic kernels for text classification using wikipedia. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 713–721. ACM.

W. Ward. 1991. Understanding spontaneous speech: the phoenix system. In *ICASSP*. IEEE.