SEMI-SUPERVISED LEARNING FOR SPOKEN LANGUAGE UNDERSTANDING USING SEMANTIC ROLE LABELING

Gokhan Tur   Dilek Hakkani-Tür

AT&T Labs - Research
Florham Park, NJ, 07932
{gtur,dtur}@research.att.com

Ananlada Chotimongkol

Carnegie Mellon University
Language Technologies Institute
Pittsburgh, PA, 15213
ananlada@cs.cmu.edu

ABSTRACT

In a goal-oriented spoken dialog system, the major aim of language understanding is to classify utterances into one or more of the pre-defined intents and extract the associated named entities. Typically, the intents are designed by a human expert according to the application domain. Furthermore, these systems are trained using large amounts of data manually labeled using an already prepared labeling guide. In this paper, we propose a semi-supervised spoken language understanding approach based on the task-independent semantic role labeling of the utterances. The goal is to extract the predicates and the associated arguments from spoken language by using semantic role labeling and determine the intents based on these predicate/argument pairs. We propose an iterative approach using the automatically labeled utterances with semantic roles as the seed training data for intent classification. We have evaluated this understanding approach using two AT&T spoken dialog system applications used for customer care. We have shown that the semantic parses obtained without using any syntactically or semantically labeled in-domain data can represent the semantic intents without a need for manual intent and labeling guide design and labeling phases. Using this approach on automatic speech recognizer transcriptions, for both applications, we have achieved the 86.5% of the performance of a classification model trained with thousands of labeled utterances.

1. INTRODUCTION

Spoken language understanding aims to extract the *meaning* of the speech utterances. In the last decade, a variety of practical goal-oriented spoken dialog systems (SDS) have been built for call routing [3, 4, 5, 6, among others]. These systems aim to identify intents of humans, expressed in natural language, and take actions accordingly, to satisfy their request. In such systems, typically, first the speaker's utterance is recognized using an automatic speech recognizer (ASR). Then, the intent (call-type) of the speaker is identified from the recognized sequence, using a spoken language understanding (SLU) component. Finally, the role of the dialog manager (DM) is to interact with the user in a natural way and help the user to achieve the task that the system is designed to support. As an example, consider the utterance *I have a question about my bill.* Assuming that the utterance is recognized correctly, the corresponding intent would be *Ask(Bill)*. Then the action that needs to be taken depends on the DM. It may ask the user to further specify the problem or route this call to the billing department.

For call-type classification, one can use a domain-dependent statistical approach as in the previous work. But this approach has some serious drawbacks. First, training statistical models for intent classification requires large amounts of labeled in-domain data, which is very expensive and time-consuming to prepare. By "labeling", we mean assigning one or more of the predefined call-type(s) to each utterance using a labeling guide. Moreover, the preparation of the labeling guide (i.e., designing the intents and the guidelines) for a given spoken language understanding task is also time-consuming and involves non-trivial design decisions. If rule-based methods are used for these tasks, this requires significant human expertise, therefore has similar problems. These decisions depend on the expert who is designing the task structure and the frequency of the intents for a given task. Furthermore, one expects the intents to be clearly defined in order to ease the job of the classifier and the human labelers. Another issue is the consistency between different tasks. This is important for manually labeling the data quickly and correctly and making the labeled data re-usable across different applications. For example in most applications, utterances like *I want to talk to a human not a machine* appear and they can be processed similarly.

On the other hand, in the computational linguistics domain, task independent semantic representations have been proposed since the last few decades. Two notable studies are FrameNet [7] and PropBank [8] projects. In this pa-

0-7803-9479-8/05/$20.00 © 2005 IEEE

per we focus on the Propbank project, which aims at creating a corpus of text annotated with information about basic semantic propositions. Predicate/argument relations are added to the syntactic trees of the existing Penn Treebank, which is mostly grammatical written text. Very recently, the PropBank corpus has been used for semantic role labeling (SRL) at the CoNLL-2004 as the shared task [9]. SRL aims to put "*who* did *what* to *whom*" kind of structures to sentences without considering the application using this information. More formally, given a predicate of the sentence, the goal of SRL is to identify all of its arguments and their semantic roles.

The relationship between the arguments of the predicates in a sentence and named entities have been previously exploited by Surdeanu *et al.* [10], who have used SRL for information extraction. In this paper, extending this idea, we propose a spoken language understanding approach based on task-independent semantic parsing of the utterances. The goal is to extract the predicates and the associated arguments from spoken language and design mapping rules to map them to some output representation which the DM can work with. This representation can be the same as or more sophisticated than the intents motivated by the possible routes in the application. We propose an iterative approach using the automatically labeled utterances (by the mapping rules) as the seed training data for intent classification. During this process *no* manual labeling or labeling guide preparation is required and the only human intervention is during the mapping rule design step, and it is miniscule compared to the traditional approach.

In the following section we explain the task of semantic role labeling in more detail. In Section 3 we present our approach of using semantic role labels for natural language understanding. Section 4 includes our experimental results using the AT&T VoiceTone® spoken dialog system data [4].

2. SEMANTIC ROLE LABELING

In the CoNLL-2004 shared task, semantic role labeling is defined as the task of analyzing the propositions expressed by some target verbs of the sentence [9]. In particular, the goal is to extract all the constituents which fill a semantic role of a target verb. Typical semantic arguments include Agent, Patient, Instrument, etc. and also adjuncts such as Locative, Temporal, Manner, Cause, etc. In the PropBank corpus, these arguments are given mnemonic names, such as Arg0, Arg1, Arg-LOC, etc. For example, for the sentence *I have bought myself a blue jacket from your summer catalog for twenty five dollars last week*, the agent (buyer, or Arg0) is *I*, the predicate is *buy*, the thing bought (Arg1) is *a blue jacket*, the seller or source (Arg2) is *from your summer catalog*, the price paid (Arg3) is *twenty five dollars*, the benefactive (Arg4) is *myself*, and the date (ArgM-TMP) is *last week*[1].

Semantic role labeling can be viewed as a multi-class classification problem. Given a word (or phrase) and its features, the goal is to output the most probable semantic role label. As it can be seen from the shared task summary paper [9], for this purpose, most researchers have used statistical classifiers with various syntactic and semantic features. The methods have ranged from Support Vector Machines (SVM) to Transformation-Based Error-Driven Learning to Memory-Based Learning. Most approaches have focused on extracting the most useful features for superior performance and have seen the classification algorithms as black boxes. PropBank corpus includes the semantic roles as well as other linguistic information, which might be useful for this task, such as part of speech tags of the words, named entities, and syntactic parses of the sentences.

In this work, we have used the exact same feature set that Hacioglu *et al.* [11] have used, since their system performed the best among others. In their approach, all features have contextual counterparts. For example the preceding and following two words, or predicted semantic roles are also used as features. Furthermore, instead of labeling the semantic role of each word, we have also employed the phrase-based labeling approach, where only the head words of phrases are labeled. This assumes that all words in a phrase have the same semantic role. Each phrase is represented with the features of the head word. This reduces the number of tokens that have to be tagged and enables the contextual features to span a larger portion of the sentence. The features include token-level features (such as the current (head) word, its part-of-speech tag, base phrase type and position, etc.), predicate-level features (such as the predicate's lemma, frequency, part-of-speech tag, etc.) and argument-level features which capture the relationship between the token (head word/phrase) and the predicate (such as the syntactic path between the token and the predicate, their distance, token position relative to the predicate, etc.).

Semantic role labeling of spoken utterances is a research challenge just by itself, because of various reasons:

- *Noisy speech recognition*: State of the art ASR systems operate with a word error rate of around 25% [4], that is they misrecognize one out of every four words. This is a big challenge for robust SRL.
- *Ungrammatical utterances with disfluencies*: Unlike the newspaper articles in the PropBank corpus, we expect the input utterances to be more casual and shorter, but on the other hand very frequently ungrammatical and including disfluencies, such as repetitions, corrections, etc.
- *Open domain*: Since the same SRL methods are going to be used for various SDS applications, such as

[1] See http://www.cis.upenn.edu/∼dgildea/Verbs for more details

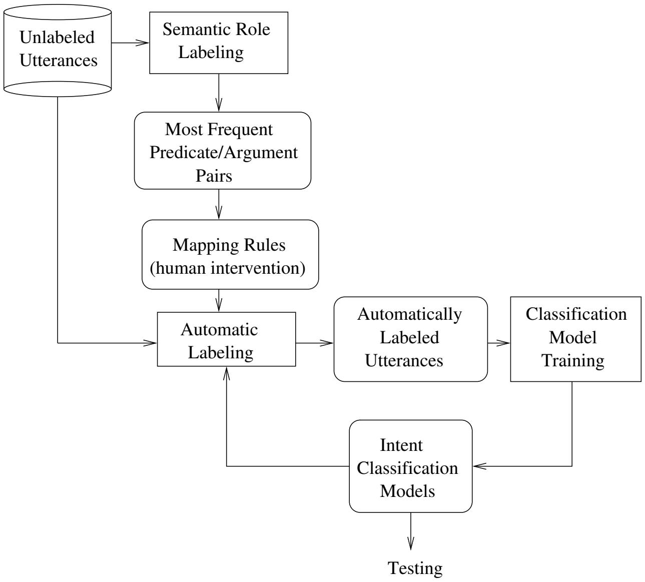
Unlabeled Utterances
Semantic Role Labeling
Most Frequent Predicate/Argument Pairs
Mapping Rules (human intervention)
Automatic Labeling
Automatically Labeled Utterances
Classification Model Training
Intent Classification Models
Testing

Fig. 1. The semi-supervised spoken language understanding approach.

> customer care systems, automobile interaction systems, etc., we expect the SRL to be robust to domain changes, and usable for many different applications with an acceptable performance.

In the CoNLL-2004 shared task, researchers have found that syntactic information, such as part of speech (POS) tags or syntactic parses and semantic information, such as named entities, are extremely useful for SRL [9]. Thus, we need to syntactically preprocess an utterance and extract named entities before semantically parsing it. This requires the feature extraction step (e.g., part of speech tagging) to face the above problems, as well.

3. APPROACH

In order to build a domain-independent spoken language understanding system, we propose using the predicates and their arguments provided by the semantic role labeling of utterances. Once an utterance is semantically parsed, we propose to extract the predicates and the related arguments and use these predicates and some certain arguments as the intents regardless of the application domain. This approach ensures the consistency across various domains and eases the job of the SDS design, which includes the determination of intents and the corresponding dialog flow. This also means that there is *no* need for in-domain data manually labeled with intents or a labeling guide to be used by human labelers. If some amount of in-domain data labeled with semantic roles is provided this would improve the performance of semantic role labeling, though it is not critical.

While building the application, the human expert is provided with the most frequent predicate/argument pairs from the training data for the domain. We use the headwords of the arguments in these pairs. The expert can then select certain predicate/argument pairs as intents by writing some *mapping rules*. For instance, consider a spoken language understanding application from a retail domain. One intent would be placing an order. For example, the utterance *I would like to place an order* would be assigned the intent *Place(Order)*. This is similar to the process of mapping a sentence to its logical form known as semantic interpretation using semantic role labels [12]. Semantically equivalent predicate/argument pairs such as *make/order* and *place/order* may be grouped while designing the intents.

One issue with this approach is caused by utterances with no predicates, such as the utterance *account balance*. Another problem is that, due to noisy ASR output, the utterance can not be parsed appropriately. In order to handle such cases we propose an iterative approach as follows: The training data is first labeled using the mapping rules. Then a statistical call-type classification model can be trained using the portion of the training data automatically labeled by the mapping rules. Using this model, the very same training data can be automatically re-labeled and the model can be re-trained, until the training set labels converge. This iterative process is depicted in Figure 1. Intuitively, using the iterative method, the statistical model can capture more features related to the call-types and hence perform better. For example, before the first round, the utterance *I'd like to know my account balance* would be labeled as *Know(Balance)* if there is such a rule for the predicate/argument pair *know/balance*. When a statistical classification model is trained with such labeled utterances, other similar utterances, such as *account balance*, may be labeled automatically with the same call-type, hence increase the amount and variability of the utterances in the training data.

4. EXPERIMENTS AND RESULTS

In this section we present the experiments and results towards a task-independent SLU. First, we present the performance of the Semantic Role Labeling system we have built using the 2004 PropBank corpus, then we present experimental results on using SRL for SLU.

4.1. PropBank Semantic Role Labeling Performance

We have trained a semantic role labeling classifier as described in Section 2 using the PropBank corpus following the CoNLL-2004 shared task. This is the Wall Street Journal part of the Penn Treebank corpus. The training set is formed from Sections 15-18, and the test set from Section 20. The number of semantic roles is 79. As the classifier

	App. 1	**App. 2**
Training Set	10,000 utt.	29,577 utt.
Test Set	5,056 utt.	5,537 utt.
No. call-types	34	96
Avg. utt. length	9.2 words	9.9 words
ASR Word Accuracy (Test)	70.3%	73.8%

Table 1. Data set characteristics

we have used Boostexter with 2000 iterations [13]. As the evaluation criteria, we have used the F-1 metric as defined in the CoNLL-2004 shared task for each semantic role (which requires both the boundary and the label to be correct) [9].

On the test set of the PropBank corpus, using a total of 113 features, we have got an F-1 value of 65.2%. Hacioglu et al. has reported an F-1 value of 71.7% on this set [9]. Our system is about 6.5% worse than theirs, though better than most of the other participants. Aside from certain implementation details, this difference might be partly due to the classifier we are using (Boostexter instead of SVM, since the training time of the former was much shorter) and the fact that we could only use less than 60% of the available training data due to memory limitations introduced by this specific implementation of Boosting.

4.2. SLU Experiments and Results

For our experiments we have used data from the retail and pharmaceutical domains, collected by the AT&T VoiceTone spoken dialog system used for customer care. Users usually call the retail system to purchase or return items, track, change, or cancel their orders, or ask about store locations, hours, etc. The other system is called mostly for refilling drugs, ordering new prescriptions, etc. It has 3 times as many call-types and training data utterances as the first one. Table 1 summarizes the characteristics of these data sets.

4.2.1. Semantic Role Labeling Performance

As the POS and NE taggers, we have used simple HMM-based taggers. In order to train the POS tagger, we have used the Penn TreeBank corpus training set. For the NE tagger we have used the MUC data [14]. We have employed Collins' parser [15], and used Buchholz's `chunklink` script to extract information from the parse trees[2].

To identify the predicates we have used a simple rule: A word is a predicate if its POS tag is a verb (except the verbs *be* and *have*, in order to be consistent with PropBank corpus). We have used a table look up to identify the predicate lemma (base form).

In order to evaluate performance of SRL on this task, we have manually annotated 285 manually transcribed utterances. They include 645 predicates (2.3 predicates/utterance). First we have computed recall and precision rates for evaluating the predicate identification performance. The precision is found to be 93.0% and recall is 91.2%. The vast majority of the errors are caused by the POS tagger, which is trained on newspaper domain. A typical example is the word *please*, which is very frequent in customer care domain but erroneously tagged as verb in most cases, since it is labeled erroneously or frequently occurs as a verb in the Penn TreeBank. More than 90% of false alarms for predicate extraction are due to this word. Most of the false rejections are due to disfluencies and ungrammatical utterances. An example would be the utterance *I'd like to order place an order*, where the predicate *place* is tagged as noun erroneously probably because of the preceding verb *order*.

Predicate	**Percent**	**Example**
place	79%	*place an order*
order	9%	*order a jacket*
make	4%	*make an order*
put	1%	*put in an order*

Table 2. Most frequent predicates for the purchase intent from a retail domain customer care application.

Then we have evaluated the argument labeling performance using a stricter measure than the CoNLL-2004 shared task. We call the labeling as correct if both the boundary and the role of all the arguments of a predicate are correct. In this work, we have ignored the mistakes on *Arg0*, since we can assume that the agent is mostly *I*, as in the utterance *checking the account balance*. In our test set, we have found out that our SRL tool correctly tags all arguments of 57.6% of the predicates. The errors are mostly due to:

- Disfluencies or sentence fragments (25%)
- Missing some arguments (25%)
- Assigning wrong argument labels (10%)
- False alarms for predicate extraction (7%)

4.2.2. Call-type Classification Performance

As the next set of experiments we have only focused on one intent, namely *Make(Order)*, from the first application, which covers utterances with purchase intents, such as *I would like to order a jacket*. In our corpus, there are 7,765 utterances with that intent (about half of all utterances). We were able to use 7,734 of them, since we could not parse the remaining 0.4% due to fragmented and cut-off sentences, or several sentences joined into one sentence. For this set of utterances, the distribution of the most frequent predicates are given in Table 2. For that call-type, one predicate (i.e., *place*) is very frequent, and there is a list of infrequent predicates.

[2] http://ilk.kub.nl/∼sabine/chunklink/chunklink_2-2-2000_for_conll.pl

Pred./Arg. pair, p	**Arg. Type**	**Call-type,** c	$P(p\|c)$	$P(c\|p)$
place/order	*Arg1*	Make(Order)	0.77	0.96
make/order	*Arg1*	Make(Order)	0.03	0.93
order/something	*Arg1*	Make(Order)	0.02	0.86
check/order	*Arg1*	Check(Order_Status)	0.14	0.95
cancel/order	*Arg1*	Cancel(Order)	0.07	0.95
check/status	*Arg1*	Check(Order_Status)	0.50	1.00
talk/someone	*Arg2*	Talk(Human)	0.05	0.89
talk/somebody	*Arg2*	Talk(Human)	0.5	0.91

Table 3. The most frequent predicate/argument pairs along with the associated call-types for the retail domain.

	App. 1				**App. 2**			
	Trans.		**ASR**		**Trans.**		**ASR**	
	R1	R2	R1	R2	R1	R2	R1	R2
Correct	56.7%	62.9%	28.1%	30.3%	42.62%	52.6%	26.3%	29.8%
No Pred/Arg	24.0%	24.0%	63.0%	63.0%	30.9%	30.9%	61.4%	61.4%
Error	3.8%	6.1%	1.7%	3.2%	6.3%	12.5%	2.8%	6.5%
No Rule	15.5%	7.0%	7.2%	3.5%	20.2%	4.0%	9.5%	2.3%

Table 4. Analysis of the call classification results using only the mapping rules using both manual and ASR transcriptions.

After these experiments, instead of considering a single call-type, we used all utterances from this application. The most frequent predicate/argument pairs are given in Table 3. For each pair, p, we compute its relation with the associated call-type, c, designed by a human user experience expert, using $P(p|c)$ and $P(c|p)$. Note that for each predicate/argument pair, there is a single call-type with a very high probability, $P(c|p)$, but a call-type may be represented by multiple pairs.

Next, we tried to perform call classification without any labeled in-domain training data. We manually grouped the most frequent predicate/argument pairs in the training data into call-types forming the mapping rules, and computed the accuracy of call classification on the test set using these. Table 4 presents the results of the call classification on the test set. We provide results using both human transcriptions and ASR outputs in order to test the robustness of our approach to noisy ASR output. We have tried using 2 mapping rule sets, R1 and R2. R2 is used for setting an upper bound with this approach where *all* predicate/argument pairs found in the training data are mapped to the most frequent call-types which have those pairs. The more realistic scenario is using R1, which consists of only the most frequent predicate/argument pairs. R1 has 80 and 170 rules and R2 has 1014 and 3396 rules for Applications 1 and 2 respectively. Some utterances had no predicate (such as *customer service please* or *account balance*) or the parser was not able to output predicate/argument pairs (as shown in No Pred/Arg row in Table 4). The other reasons for classification mistakes are incorrect mapping rules (Error) and absence of mapping rules from predicate/argument pairs to calltypes (No Rule). The absence of a mapping rule was mainly caused by data sparseness and the absence of argument grouping. For example, even though the pair *order/pants* was in the training data, *order/trousers* was not. As can be seen from both this table, the performances on ASR transcriptions using these mapping rules are pretty low, mostly due to the lack of robustness of the semantic parser for the ASR errors.

Finally, we employed the proposed iterative approach. The results are provided in Table 5. Even with one iteration, there is a significant jump in the performance, especially for the ASR, since the model has become more robust to ASR errors. With the upper-bound experiment, using an extensive mapping rule set, we achieved around 90% (e.g. 79.7% instead of 88.7%) of the performance to that of the supervised model. Using only a small rule set, this number reduces to only 86.5% on ASR transcriptions for both applications.

5. CONCLUSIONS AND FUTURE WORK

We have presented a semi-supervised spoken language understanding approach depending on the semantic role labels in an utterance. We have demonstrated the use of this approach using two real-life SDS applications from retail and pharmaceutical domains. Using a small rule set, with *no* labeled in-domain data, using both ASR output and human transcriptions, for both applications, we have achieved the 86.5% of the performance of a model trained with thousands of labeled utterances. We have seen that with manual transcriptions, ungrammatical fragments and disfluencies cause less problem than expected although the semantic role la-

Iteration	App. 1				App. 2			
	Trans.		ASR		Trans.		ASR	
	R1	R2	R1	R2	R1	R2	R1	R2
0	56.7%	62.9%	28.1%	30.3%	42.6%	52.6%	26.3%	29.8%
1	**76.6%**	**79.7%**	71.1%	**75.7%**	66.8%	**70.7%**	63.4%	**66.3%**
2	74.2%	78.3%	**71.5%**	74.3%	67.4%	70.5%	64.2%	66.2%
3	74.0%	-	71.5%	-	**67.6%**	-	**64.4%**	-
SUPERVISED	88.7%	88.7%	82.7%	82.7%	81.8%	81.8%	74.4%	74.4%

Table 5. Call classification results for the iterative approach using both manual and ASR transcriptions with different rule sets. The best performance for each case is marked with boldface.

beling tool and the underlying part of speech tagger, named entity extractor, and syntactic parser are trained using textual data, mostly newspaper articles. We have seen that SRL is good at handling the variation in input sentences. This is mostly due to the fact that the utterances we deal with are generally short and simple to process. Although semantic parsing suffered from the ASR errors, the iterative approach greatly eliminated this drawback.

Note that, the approach is expected to perform inferior on infrequent call-types, which can be captured using an active learning approach [16]. Furthermore, the verbs *be* and *have* are not marked as predicates in the PropBank corpus. This causes utterances such as *I have a billing question* to have no predicate. For our SLU approach, we would like to have these verbs as special predicates in order to distinguish them from utterances which do not have a predicate.

As future work, we plan to improve the semantic role labeler, especially using some labeled spoken dialog data, and experiment with data from new domains. We also plan to enhance input pre-processing module for cases such as disfluencies. Another research direction is exploring tighter integration of ASR and SLU for semantic role labeling and necessary feature extraction such as part of speech tagging, syntactic parsing, and named entity extraction. For example, ASR output may be more than just the 1-best string (best hypothesis), and also include multiple hypotheses, as well as word confidence scores, which provide an estimate of the correctness of the recognized words.

6. REFERENCES

[1] R. Jackendoff, *Foundations of Language*, chapter 9, Oxford University Press, 2002.

[2] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 49, no. 236, pp. 433–460, 1950.

[3] A. L. Gorin, G. Riccardi, and J. H. Wright, "How May I Help You?," *Speech Communication*, vol. 23, pp. 113–127, 1997.

[4] N. Gupta, G. Tur, D. Hakkani-Tür, S. Bangalore, G. Riccardi, and M. Rahim, "The AT&T spoken language understanding system," *IEEE Transactions on Speech and Audio Processing*, To appear.

[5] J. Chu-Carroll and B. Carpenter, "Vector-based natural language call routing," *Computational Linguistics*, vol. 25, no. 3, pp. 361–388, 1999.

[6] P. Natarajan, R. Prasad, B. Suhm, and D. McCarthy, "Speech enabled natural language call routing: BBN call director," in *Proceedings of the ICSLP*, Denver, CO, September 2002.

[7] C. J. Fillmore J. B. Lowe, C. F. Baker, "A frame-semantic approach to semantic annotation," in *Proceedings of the ACL - SIGLEX Workshop*, Washington, D.C., April 1997.

[8] P. Kingsbury, M. Marcus, and M. Palmer, "Adding semantic annotation to the Penn TreeBank," in *Proceedings of the HLT*, San Diego, CA, March 2002.

[9] X. Carreras and L. Màrquez, "Introduction to the CoNLL-2004 shared task: Semantic role labeling," in *Proceedings of the CoNLL*, Boston, MA, May 2004.

[10] M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth, "Using predicate-argument structures for information extraction," in *Proceedings of the ACL*, Sapporo, Japan, July 2003.

[11] K. Hacioglu, S. Pradhan, W. Ward, J. H. Martin, and D. Jurafsky, "Semantic role labeling by tagging syntactic chunks," in *Proceedings of the CoNLL*, Boston, MA, May 2004.

[12] J. Allen, *Natural Language Understanding*, chapter 8, Benjamin/Cummings, 1995.

[13] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000.

[14] *Proceedings of the 7^{th} Message Understanding Conference (MUC-7)*, Fairfax, VA, April 1998.

[15] M. Collins, *Head-Driven Statistical Models for Natural Language Parsing*, Ph.D. thesis, University of Pennsylvania, 1999.

[16] G. Tur, R. E. Schapire, and D. Hakkani-Tür, "Active learning for spoken language understanding," in *Proceedings of the ICASSP*, Hong Kong, May 2003.