WHAT IS LEFT TO BE UNDERSTOOD IN ATIS?

Gokhan Tur   Dilek Hakkani-Tür   Larry Heck

Speech at Microsoft | Microsoft Research
Mountain View, CA, 94041
gokhan.tur@ieee.org dilek@ieee.org larry.heck@microsoft.com

ABSTRACT

One of the main data resources used in many studies over the past two decades for spoken language understanding (SLU) research in spoken dialog systems is the airline travel information system (ATIS) corpus. Two primary tasks in SLU are intent determination (ID) and slot filling (SF). Recent studies reported error rates below 5% for both of these tasks employing discriminative machine learning techniques with the ATIS test set. While these low error rates may suggest that this task is close to being solved, further analysis reveals the continued utility of ATIS as a research corpus. In this paper, our goal is not experimenting with domain specific techniques or features which can help with the remaining SLU errors, but instead exploring methods to realize this utility via extensive error analysis. We conclude that even with such low error rates, ATIS test set still includes many unseen example categories and sequences, hence requires more data. Better yet, new annotated larger data sets from more complex tasks with realistic utterances can avoid over-tuning in terms of modeling and feature design. We believe that advancements in SLU can be achieved by having more naturally spoken data sets and employing more linguistically motivated features while preserving robustness due to speech recognition noise and variance due to natural language.

***Index Terms*—** spoken language understanding, ATIS, discriminative training

1. INTRODUCTION

Spoken language understanding (SLU) aims to extract the *meaning* of the speech utterances. While understanding language is still considered an unsolved problem, in the last decade, a variety of practical goal-oriented conversational understanding systems have been built for limited domains. These systems aim to automatically identify the intent of the user as expressed in natural language, extract associated arguments or slots, and take actions accordingly to satisfy the user's requests. In such systems, the speaker's utterance is typically recognized using an automatic speech recognizer (ASR). Then the intent of the speaker is identified from the recognized word sequence using an SLU component. Finally, a dialog or task manager (DM) interacts with the user (not necessarily in natural language) and helps the user achieve the task that the system is designed to support.

In the early 90s, DARPA (Defense Advanced Research Program Agency) initiated the Airline Travel Information System (ATIS) project. The ATIS task consisted of spoken queries on flight-related information. An example utterance is *I want to fly to Boston from New York next week*. Understanding was reduced to the problem of extracting task-specific arguments, such as *Destination* and *Departure Date*. Participating systems employed either a data-driven statistical approach [1, 2] or a knowledge-based approach [3, 4, 5].

Almost simultaneously with the semantic frame filling-based SLU approaches, a new task emerged motivated by the success of the early commercial interactive voice response (IVR) applications used in call centers. The SLU was framed as *classifying users' utterances into predefined categories* (called as *intents* or *call-types*) [6].

The biggest difference between the call classification systems and semantic frame filling systems is that the former does not explicitly seek to determine the arguments provided by the user. The main goal is *routing* the call to an appropriate call center department. The arguments provided by the user are important only in the sense that they help make the right classification. While this has been a totally different perspective for the task of SLU, it was actually complementary to template filling in that each call-type can be viewed as a template to be filled. For example, in the case of the DARPA ATIS project, while the primary *intent* (or goal) was *Flight*, users also asked about many other things such as *Ground transportation* or *Airplane specifications*. The program also defined specialized templates for these less frequent intents. This led to a seamless integration of intent determination (ID) and slot filling (SF) based SLU approaches. This integrated approach actually yielded *improved* end-to-end automation rates as compared to the previous decoupled and sequential approaches. For example, Jeong *et al* [7] proposed to model these two systems *jointly* using a triangular chain conditional random field (CRF).

In this paper, rather than focus on specific techniques or features to improve ID and SF accuracy, our goal is to assess the continued utility of the ATIS corpus given the two decades of research it has supported. In the next section, we briefly describe the ATIS corpus and then discuss the evaluation metrics for ID and SF. In Section 4, we present the state-of-the-art discriminative training efforts for both ID and SF for the task of ATIS. Finally, in Sections 5 and 6 we present our detailed analyses on the errors we have seen using ID and SF models, respectively, with comparable performance to those reported in the literature. We will show that, by categorizing the erroneous cases that remain after N-fold cross validation experiments, ATIS is still useful and suggests future research directions in SLU.

2. AIRLINE TRAVEL INFORMATION (ATIS) CORPUS

An important by-product of the DARPA ATIS project was the ATIS corpus. This corpus is the most commonly used data set for SLU research [8]. The corpus has seventeen different intents, such as *Flight* or *Aircraft capacity*. The prior distribution is, however, heavily skewed, and the most frequent intent, *Flight* represents about 70% of the traffic. Table 1 shows the frequency of the intents in this corpus for training and test sets.

In this paper, we use the ATIS corpus as used in He and Young [9] and Raymond and Riccardi [10]. The training set contains 4,978 ut-

978-1-4244-7903-0/10/$26.00 ©2010 IEEE

Intent	Training Set	Test Set
Abbreviation	2.4%	3.6%
Aircraft	1.6%	0.9%
Airfare	9.0%	5.8%
Airline	3.4%	4.3%
Airport	0.5%	2.0%
Capacity	0.4%	2.4%
City	0.3%	0.6%
Day_Name	0.1%	0.1%
Distance	0.4%	1.1%
Flight	73.1%	71.6%
Flight_No	0.3%	1.0%
Flight_Time	1.2%	0.1%
Ground_Fare	0.4%	0.8%
Ground_Service	5.5%	4.0%
Meal	0.1%	0.6%
Quantity	1.1%	0.9%
Restriction	0.3%	0.1%

Table 1. The frequency of intents for the training and test sets.

Utterance	*How much is the cheapest flight from Boston to New York tomorrow morning?*
Goal:	Airfare
Cost_Relative	*cheapest*
Depart_City	*Boston*
Arrival_City	*New York*
Depart_Date.Relative	*tomorrow*
Depart_Time.Period	*morning*

Table 2. An example utterance from the ATIS dataset.

terances selected from the Class A (context independent) training data in the ATIS-2 and ATIS-3 corpora, while the test set contains 893 utterances from the ATIS-3 Nov93 and Dec94 datasets. Each utterance has its named entities marked via table lookup, including domain specific entities such as city, airline, airport names, and dates.

The ATIS utterances are represented using semantic frames, where each sentence has a goal or goals (a.k.a. intent) and slots filled with phrases. The values of the slots are not normalized or interpreted. An example utterance with annotations is shown in Table 2.

3. EVALUATION METRICS

The most commonly used metrics for ID and SF are class (or slot) error rate (ER) and F-Measure. The simpler metric ER for ID can be computed as:

$$ER_{ID} = \frac{\text{\# misclassified utterances}}{\text{\# utterances}}$$

Note that one utterance can have more than one intent. A typical example is *Can you tell me my balance? I need to make a transfer.* In most cases, where the second intent is generic (a greeting, small talk with the human agent) or vague, it is ignored. If none of the true classes is selected, it is counted as a misclassification.

For SF, the error rate can be computed in two ways: The more common metric is the F-measure using the slots as units. This metric is similar to what is being used for other sequence classification tasks in the natural language processing community, such as parsing and named entity extraction. In this technique, usually the IOB schema is adopted, where each of the words are tagged with their position in the slot: beginning (B), in (I) or other (O). Then, recall and precision values are computed for each of the slots. A slot is considered to be correct if its range and type are correct. The F-Measure is defined as the harmonic mean of recall and precision:

$$F - Measure = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

where

$$Recall = \frac{\text{\# correct slots found}}{\text{\# true slots}}$$

$$Precision = \frac{\text{\# correct slots found}}{\text{\# found slots}}$$

.

4. BACKGROUND ON USING DISCRIMINATIVE CLASSIFIERS FOR SLU

With advances in machine learning over the last decade, especially in discriminative classification techniques, researchers have framed the ID problem as a sample classification task and SF as a sequence classification task. Typically, word n-grams are used as features after preprocessing with generic entities, such as dates, locations, or phone numbers. Because of the very large dimension of the input space, large margin classifiers such as SVMs [11] or Adaboost[12] were found to be very good candidates for ID and CRFs [13] for SF. To take into account context, the recent trend is to match n*-grams* (a substring of n words) rather than words.

As discovered, data driven approaches are very well-suited for processing spontaneous spoken utterances. The data driven approaches are typically more robust to sentences that are not well-formed grammatically, which occurs frequently in spontaneous speech. Even in broadcast conversations where participants are very well trained and prepared, a large percentage of the utterances have disfluencies: repetitions, false starts, and filler words (e.g., *uh*) [14]. Furthermore, speech recognition introduces significant "noise" to the SLU component caused by background noise, mismatched domains, incorrect recognition of proper names (such as city or person names), and reduced accuracy due to sub-real time processing requirements. A typical call routing system operates at around 20%-30% word error rate; one out of every three to five words is wrong [15]. Given that the researchers in this study also determined that one third of the ID errors are due to speech recognition noise, robust methods for spontaneous speech recognition are critically important for successful ID and SF in SLU systems. To this end, researchers have proposed many methods ranging from N-best rescoring, exploiting word confusion networks, and leveraging dialog context as prior knowledge (e.g., [15]).

4.1. Intent Determination

For ID, early work with discriminative classification algorithms was completed on the AT&T HMIHY system [6] using the Boostexter tool, an implementation of the AdaBoost.MH multiclass multilabel classification algorithm [12]. Hakkani-Tür *et al.* extended this work by using a lattice of syntactic and semantic features [16]. Discriminative call classification systems employing large margin classifiers (e.g., support vector machines) include work by Haffner *et al.* [17], who proposed a global optimization process based on an optimal

Correct-Estimated	a	b	b	d	e	f	g	h	i	j	k	l	m	n	o	p	q
a. *Abbreviation*	30									2							
b. *Aircraft*		6								3							
c. *Airfare*			64							1							
d. *Airline*				37						2							
e. *Airport*					15					2					1		
f. *Capacity*	1	5				13				2							
g. *City*							3			2							
h. *Day_Name*										2							
i. *Distance*									9	1							
j. *Flight*		1	1				1			623							
k. *Flight_No*										2	6						
l. *Flight_Time*										1							
m. *Ground_Fare*			1										3	3			
n. *Ground_Service*														36			
o. *Meal*										5							
p. *Quantity*																8	
q. *Restriction*	1																

Table 3. The confusion matrix for intent determination.

channel communication model that allowed a combination of *heterogeneous* binary classifiers. This approach decreased the call-type classification error rate for AT&T's HMIHY natural dialog system significantly, especially the false rejection rates.

Other work by Kuo and Lee [18] at Bell Labs proposed the use of discriminative training on the routing matrix, significantly improving their vector-based call routing system [19] for low rejection rates. Their approach is based on using the minimum classification error (MCE) criterion. Later they extended this approach to include Boosting and automatic relevance feedback (ARF) [20]. Cox [21] proposed the use of generalized probabilistic descent (GPD), corrective training (CT), and linear discriminant analysis (LDA).

Finally, Chelba *et al.* proposed using Maximum Entropy models for ID, and compared the performance with a Naive Bayes approach with the ATIS corpus. The discriminative method resulted in half the classification error rate compared to Naive Bayes on this highly skewed data set. They have reported about 4.8% top class error rate using a slightly different training and test corpora than the one used in this paper.

4.2. Slot Filling

For SF, the ATIS corpus has been extensively studied from the early days of the DARPA ATIS project. However, the use of discriminative classification algorithms is more recent. Some notable studies include the following:

Wang and Acero [22] compared the use of CRF, perceptron, large margin, and MCE using stochastic gradient descent (SGD) for SF in the ATIS domain. They obtained significantly reduced slot error rates, with best performance achieved by CRF (though it was the slowest to train).

Almost simultaneously Jeong and Lee [7] proposed the use of CRF, extended by non-local features, which are important to disambiguate the type of the slot. For example, a day can be the arrival day, departure day, or the return day. If the contextual cues disambiguating them are beyond the immediate context, it is not easy for the classifier to choose the correct class. Using non-local trigger features automatically extracted from the training data is shown to improve the performance significantly.

Finally, Raymond and Riccardi [10] compared SVM and CRF with generative models for the ATIS task. They concluded that discriminative methods perform significantly better, and furthermore, it is possible to incorporate a-priori information or long distance features easily. For example they added features such as "Does this utterance have the verb *arrive*". This resulted in about 10% relative reduction in slot error rate. The design of such features usually requires domain knowledge.

5. ANALYSIS OF INTENT DETERMINATION IN ATIS

In this section, our goal is to analyze the errors of a state-of-the-art ID system for the ATIS domain, cluster the errors, and then categorize the error types. These categories of error types will suggest potential areas of research that could yield improved accuracy. All experiments and analyses are performed using manual transcriptions of the training and test sets to isolate the study from noise introduced by the speech recognizer.

5.1. Discriminative Training and Experiments

For the following experiments, we used the ATIS corpus as described previously in Section 2. Since the superior performance of the discriminative training algorithms has been shown by the earlier work, we have employed the AdaBoost.MH algorithm in this study. We used only word n-grams as features. We have not optimized Boosting parameters on a tuning set nor learned weak classifiers. The data is normalized to lower case, but no stemming or stopword removal has been performed.

The ATIS test set was classified according to the classes defined in Table 1. The ID error rate we obtained was 4.5%, which is comparable to (and actually lower than) to what has been reported in the literature.

5.2. Analysis of Intent Determination Errors

Next, we checked the ID errors with three training and test set-ups:

1. *All Train*: uses all ATIS training data to train the model, and errors are computed on the ATIS test set. In total, this model erroneously classified only 40 utterances (an error rate of 4.5%). The intent confusion matrix for these errors is provided in Table 3.

2. *25% Train*: uses 25% of the training examples in the ATIS training set, and errors are computed on the ATIS test set. In total, this model erroneously classified 65 utterances (an error rate of 7.3%).

3. *N-fold*: uses all examples for both testing and training in 10-fold cross validation experiments. In total, this model erroneously classified 162 utterances (an error rate of 3.0%).

As seen in Table 3, the problem is mostly the non-*Flight* utterances erroneously classified as *Flight*. While one cause of these errors is the unbalanced intent distribution, we have manually checked each error and clustered them into 6 categories:

1. *Prepositional phrases embedded in noun phrases*: These errors involve phrases such as *Capacity of the flight from Boston to Orlando*, where the prepositional phrase suggests flight information, whereas the destination category is mainly determined by the head word of the noun phrase (*capacity* in this case). Since classifier has no syntactic features, such sentences are usually classified erroneously. Using features from a syntactic parser can alleviate this problem.

2. *Wrong functional arguments of utterances*: This category is similar to the first category but the difference is that, instead of a prepositional phrase, the confused phrase is a semantic argument of the utterance. Consider the example utterance *What day of the week does the flight from Boston to Orlando fly?* These are errors that can be solved by using either a syntactic parser that identifies functions of phrases or a semantic role labeler.

3. *Annotation errors*: These are utterances that were assigned the wrong category during manual annotation.

4. *Utterances with multiple sentences*: These are utterances with more than one sentence. In such cases, the intent is usually in the last sentence, whereas the classification output is biased by the other sentence.

5. *Other*: These include several infrequent error types such as ambiguous utterances, ill-formulated queries, and preprocessing/tokenization issues:

   - *Ambiguous utterances*: These errors involve utterances where the destination category is not clear in the utterance. An example from the ATIS test set is *list Los Angeles*. In this utterance, the speaker intent could either be to find cities that have flights *from* Los Angeles or flights *to* Los Angeles.

   - *Ill-formulated queries*: These are utterances which include a phrase that may mislead the classification or understanding. An example from the ATIS test set is: *What's the airfare for a taxi to the Denver airport?* In this case, the word *airfare* implies a destination category of *Airfare*, whereas what is meant is *Ground transportation fare*. These type of errors are easier for humans to handle, but it is not presently clear how they can be resolved in automatic processing.

   - *Preprocessing/Tokenization issues*: These are errors that could be resolved by using a domain ontology or special pre-processing or tokenization related to the domain. Some domain specific abbreviations and restriction codes are examples of this category.

Error Type	All Train	25% Train	10-Fold
1	42.5%	33.8%	24.5%
2	22.5%	13.8%	30.0%
3	2.5%	6.1%	18.4%
4	0%	0%	8.0%
5	17.5%	12.5%	7.2%
6	15.0%	33.8%	11.7%

Table 4. The distribution of error categories for ID using all and 25% of the training data, and using all the training and the test set with 10-fold cross validation.

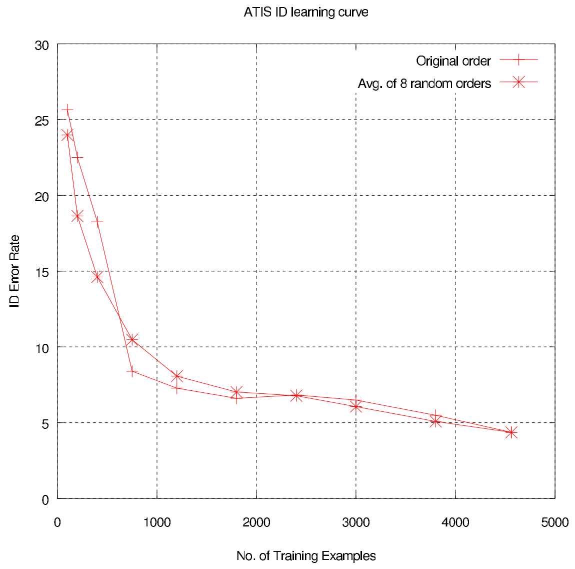

Fig. 1. Learning curve for intent determination using the training data with the original order and average of 8 shuffled orders.

6. *Difficult Cases*: These are utterances that include words or phrases that were previously unseen in the training data. For the example utterance *Are snack served on Tower Air?*, none of the content words and phrases appear with the *Meal* category in the training data.

Table 4 presents the frequency of each of these errors for the three experiments. As seen, categories 1 and 2 constitute a majority of the errors. Both of these categories can be resolved using a syntactic parser with function tags. However, note that the ATIS corpus is highly artificial and utterances are mostly grammatical and without disfluencies. Furthermore, when working with ASR, utterances may include recognition errors. In a more realistic scenario, one might consider shallow parsing or syntactic and semantic graphs [16] for extracting richer and linguistically-motivated features that could resolve such cases.

Figure 1 shows the error rate on the ATIS test set when varying training set sizes are used. When manually examining the test set, we found clusters of similar utterances occurring one after the other (probably are uttered by the same user). To eliminate the bias from the data collection order, we also estimated the error with a random ordering of the training set, and averaged the error rates over 8 such experiments. As can be seen from this plot, the error rate keeps shrinking as more data is added, suggesting that more training data would be beneficial.

6. ANALYSIS OF SLOT FILLING IN ATIS

In this section, our goal is similar to the ID analysis: analyze the results of a state-of-the-art SF system for the ATIS domain and cluster the errors into categories.

6.1. Discriminative Training and Experiments

Following methods described in the literature, we employed linear chain CRFs to model the slots in the ATIS Domain. We used only word n-gram features and did not use a development set to tune parameters. The ATIS test set was then classified using the trained model. We converted the data sets into the IOB format so that we have only one word per sample to classify. Using the CoNLL evaluation script[1], the SF F-Measure we obtained was 93.2% with the IOB representation[2], which is comparable to what has been reported in the literature.

6.2. Analysis of Slot Filling Errors

Analyzing the SF decisions, the model found 2,614 of 2,837 slots with the correct type and span for the input out of 9,164 words. We manually checked each of the 223 erroneous cases and clustered them into 8 categories:

1. *Long distance dependencies*: These are slots where the disambiguating tokens are out of the current n-gram context. For example, in the utterance *Find flights to New York arriving in no later than next Saturday*, a 6-gram context is required to resolve that *Saturday* is the arrival date. This category was previously addressed in the literature. For example, Raymond and Riccardi [10] extracted features using manually-designed patterns and Jeong and Lee [7] used trigger patterns to cover these cases.
2. *Partially correct slot value annotations*: These are slots assigned a category that is partially correct; either the category or the sub-category matches the manual annotation. For example, the word *tomorrow* can either be a *Depart_Date.Relative* or *Arrive_Date.Relative* for the utterance *flights arriving in Boston tomorrow*. Note that these can overlap with other error types.
3. *Previously unseen sequences:* While this category requires further analysis, the most common reason is the mismatch between the training and test sets. For example, *meal* related slots are missed by the model (8.0% of all errors) because there are no similar cases in the training set. This is also the case for the aircraft models (10.0%), and traveling to states instead of cities (3.3%), etc.
4. *Annotation errors*: These are the slots that were assigned the wrong category during manual annotation.
5. *Other*: These include several infrequent error types such as ambiguous utterances, ill-formulated queries, and preprocessing/tokenization issues:
   - *Ill-formulated queries*: These errors usually involve an ungrammatical phrase that may mislead the interpretation of the slot value or there is insufficient context to disambiguate the value of the slot. For example, in the utterance *Find a flight from Memphis to Tacoma dinner*, it is not clear if the word *dinner* refers to the description of the flight meal.
   - *Ambiguous utterances*: These are utterances where the slot category is not explicit given the utterance. For example, in the utterance *I would like to have the airline that flies Toronto, Detroit and Orlando*, it is not clear if the speaker is searching for airlines that have flights from Toronto to Detroit and Orlando or from some other location to Toronto, Detroit and Orlando.
   - *Preprocessing/Tokenization issues*: These are errors that could be resolved using a domain ontology or special pre-processing or tokenization related to the domain. For example, in the utterance *What airline is AS*, it would be helpful to know *AS* is a domain specific abbreviation.
   - *Ambiguous part-of-speech tag-related errors:* These are errors that could be resolved if the part-of-speech tags were resolved. For example, the word *arriving* can be a verb or an adjective, as in the utterance *I want to find the earliest arriving flight to Boston*. In this case, the slot category for the words *earliest arriving* is *Flight-Mod*, but since the word *arriving* is very frequently seen as a verb in this corpus, it is assigned no slot category.

Error Type	Percentage
1	26.9%
2	42.4%
3	57.6%
4	8.4%
5	6.7%

Table 5. The distribution of the types of errors in the ATIS test set. Note that these do not sum to 100% as some errors include multiple types.

Table 5 lists the frequency of each of these errors. Categories 1, 2, and 3 constitute vast majority of the errors. Each of these categories can be attacked using a different strategy. Category 1 utterances are the easiest to resolve using richer feature sets during discriminative training. Using a-priori information may also help when available. Also, discovering linguistically motivated long distance patterns is a promising research work. Category 2 utterances happen mainly due to the nuance between the *arrive* and *depart* concepts (23.1% of all errors), which are very hard to distinguish in some cases as in the example above. Category 3 utterances simply require a better training set or human intervention of manual patterns as they are underrepresented or missing in the training data.

7. DISCUSSION AND CONCLUSIONS

Leveraging recent improvements in machine learning and spoken language processing, the performance of the SLU systems for the ATIS domain has improved dramatically. Around 5% error rate for the SLU task implies a solved problem. It is clear, however, that the problem of SLU is far from being solved, especially for more realistic, naturally-spoken utterances of a variety of speakers from tasks more complex than simple flight information requests. New data sets from such tasks can avoid over-tuning to one particular data set in terms of modeling and feature design.

[1] http://www.cnts.ua.ac.be/conll2000/chunking/output.html

[2] It is 94.7% using the representation used by [10], who reported 95.0%

The recent French Media corpus [23] offers a step towards this goal: it has three times more data and greater than a 10% concept error rate for SF. However, the data was not collected from an operational system. Instead, data was collected using a wizard of Oz setup with selected volunteers. Another effort is the Let's Go dialog system used by real users of the Pittsburgh bus transportation system [24]. However, SLU annotations are not yet available.

Even with such low error rates, the ATIS test set includes many example categories and sequences unseen in the training data, and the error rates have not converged yet. In that respect, more data from just the ATIS domain may be useful for SLU research.

The error analysis on the ATIS domain shows the primary weaknesses of the current n-gram-based modeling approaches: The local context overrides the global, the model has no domain knowledge to make any inferences, and it tries to fit any utterance into some known sample, hence not really robust to any out-of-domain utterances. This was also observed by Raymond and Riccardi [10], where the CRF model fits 100% to the training data. One possible research direction consists of employing longer distance syntactically or semantically motivated features, while preserving the robustness of the system to the noise introduced by the speech recognizer and variance due to natural language.

A lesser studied set of the ATIS corpus, Class D utterances, which are contextual queries, is another significant portion of this corpus, waiting to be understood. While most people treated understanding in context with handcrafted rules (e.g., [4]), to the best of our knowledge, the only study towards building a statistical discourse model has been proposed by Miller *et al.* [25].

8. ACKNOWLEDGMENTS

We would like to thank Christian Raymond and Giuseppe Riccardi for sharing the ATIS data, revised for annotation inconsistencies and mistakes.

9. REFERENCES

[1] R. Pieraccini, E. Tzoukermann, Z. Gorelov, J.-L. Gauvain, E. Levin, C.-H. Lee, and J. G. Wilpon, "A speech understanding system based on statistical representation of semantics," in *Proceedings of the ICASSP*, San Francisco, CA, March 1992.

[2] S. Miller, R. Bobrow, R. Ingria, and R. Schwartz, "Hidden understanding models of natural language," in *Proceedings of the ACL*, Las Cruces, NM, June 1994.

[3] W. Ward and S.Issar, "Recent improvements in the CMU spoken language understanding system," in *Proceedings of the ARPA HLT Workshop*, March 1994, pp. 213–216.

[4] S. Seneff, "TINA: A natural language system for spoken language applications," *Computational Linguistics*, vol. 18, no. 1, pp. 61–86, 1992.

[5] J. Dowding, J. M. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore, and D. Moran, "Gemini: A natural language system for spoken language understanding," in *Proceedings of the ARPA Workshop on Human Language Technology*, Princeton, NJ, March 1993.

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

[7] M. Jeong and G. G. Lee, "Exploiting non-local features for spoken language understanding," in *Proceedings of the ACL/COLING*, Sydney, Australia, July 2006.

[8] P. J. Price, "Evaluation of spoken language systems: The ATIS domain," in *Proceedings of the DARPA Workshop on Speech and Natural Language*, Hidden Valley, PA, June 1990.

[9] Y. He and S. Young, "A data-driven spoken language understanding system," in *Proceedings of the IEEE ASRU Workshop*, U.S. Virgin Islands, December 2003, pp. 583–588.

[10] C. Raymond and G. Riccardi, "Generative and discriminative algorithms for spoken language understanding," in *Proceedings of the Interspeech*, Antwerp, Belgium, 2007.

[11] V. N. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, New York, NY, 1998.

[12] 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.

[13] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the ICML*, Williamstown, MA, 2001.

[14] A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies," in *Proceedings of the ICASSP*, Atlanta, GA, May 1996.

[15] N. Gupta, G. Tur, D. Hakkani-Tür, S. Bangalore, G. Riccardi, and M. Rahim, "The AT&T spoken language understanding system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 213–222, 2006.

[16] D. Hakkani-Tür, G. Tur, and A. Chotimongkol, "Using syntactic and semantic graphs for call classification," in *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, Ann Arbor, MI, June 2005.

[17] P. Haffner, G. Tur, and J. Wright, "Optimizing SVMs for complex call classification," in *Proceedings of the ICASSP*, Hong Kong, April 2003.

[18] H.-K. J. Kuo and C.-H. Lee, "Discriminative training in natural language call-routing," in *Proceedings of ICSLP*, Beijing, China, 2000.

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

[20] I. Zitouni, H.-K. J. Kuo, and C.-H. Lee, "Boosting and combincation of classifiers for natural language call routing systems," *Speech Communication*, vol. 41, no. 4, pp. 647–661, 2003.

[21] Stephen Cox, "Discriminative techniques in call routing," in *Proceedings of the ICASSP*, Hong Kong, April 2003.

[22] Y.-Y. Wang and A. Acero, "Discriminative models for spoken language understanding," in *Proceedings of the ICSLP*, Pittsburgh, PA, September 2006.

[23] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa, "Semantic annotation of the French MEDIA dialog corpus," in *Proceedings of the Interspeech*, Lisbon, Portugal, September 2005.

[24] A. Raux, B. Langner, D. Bohus, A. Black, and M. Eskenazi, "Let's go public! taking a spoken dialog system to the real world," in *Proceedings of the Interspeech*, Lisbon, Portugal, September 2005.

[25] S. Miller, D. Stallard, R. Bobrow, and R. Schwartz, "A fully statistical approach to natural language interfaces," in *Proceedings of the ACL*, Morristown, NJ, 1996.