Experiments on the France Telecom 3000 Voice Agency corpus: academic research on an industrial spoken dialog system*

Géraldine Damnati
France Télécom R&D
TECH/SSTP/RVA
2 av. Pierre Marzin
22307 Lannion Cedex 07, France
geraldine.damnati@orange-ftgroup.com

Frédéric Béchet **Renato De Mori**
LIA
University of Avignon
AGROPARC, 339 ch. des Meinajaries
84911 Avignon Cedex 09, France
frederic.bechet,renato.demori
@univ-avignon.fr

Abstract

The recent advances in speech recognition technologies, and the experience acquired in the development of WEB or Interactive Voice Response interfaces, have facilitated the integration of speech modules in robust Spoken Dialog Systems (SDS), leading to the deployment on a large scale of speech-enabled services. With these services it is possible to obtain very large corpora of human-machine interactions by collecting system logs. This new kinds of systems and dialogue corpora offer new opportunities for academic research while raising two issues: How can academic research take profit of the system logs of deployed SDS in order to build the *next generation* of SDS, although the dialogues collected have a dialogue flow constrained by the *previous SDS generation*? On the other side, what immediate benefits can academic research offer for the improvement of deployed system? This paper addresses these aspects in the framework of the deployed France Telecom 3000 Voice Agency service.

*This work is supported by the 6th Framework Research Programme of the European Union (EU), Project LUNA, IST contract no 33549. The authors would like to thank the EU for the financial support. For more information about the LUNA project, please visit the project home-page, www.ist-luna.eu .

1 Introduction

Since the deployment on a very large scale of the AT&T *How May I Help You?* (HMIHY) (Gorin et al., 1997) service in 2000, Spoken Dialogue Systems (SDS) handling a very large number of calls are now developed from an industrial point of view. Although a lot of the remaining problems (robustness, coverage, etc.) are still spoken language processing research problems, the conception and the deployment of such state-of-the-art systems mainly requires knowledge in user interfaces.

The recent advances in speech recognition technologies, and the experience acquired in the development of WEB or Interactive Voice Response interfaces have facilitated the integration of speech modules in robust SDS.

These new SDS can be deployed on a very large scale, like the France Telecom 3000 Voice Agency service considered in this study. With these services it is possible to obtain very large corpora of human-machine interactions by collecting system logs. The main differences between these corpora and those collected in the framework of evaluation programs like the DARPA ATIS (Hemphill et al., 1990) or the French Technolangue MEDIA (Bonneau-Maynard et al., 2005) programs can be expressed through the following dimensions:

- Size. There are virtually no limits in the amount of speakers available or the time needed for collecting the dialogues as thousands of dialogues are automatically processed every day and the system logs are stored. Therefore Dialog processing becomes similar

Bridging the Gap: Academic and Industrial Research in Dialog Technologies Workshop Proceedings, pages 48–55,
NAACL-HLT, Rochester, NY, April 2007.

to Broadcast News processing: the limit is not in the amount of data available, but rather in the amount of data that can be manually annotated.

- Speakers. Data are from *real* users. The speakers are not professional ones or have no reward for calling the system. Therefore their behaviors are not biased by the acquisition protocols. Spontaneous speech and speech affects can be observed.

- Complexity. The complexity of the services widely deployed is necessarily limited in order to guarantee robustness with a high automation rate. Therefore the dialogues collected are often short dialogues.

- Semantic model. The semantic model of such deployed system is task-oriented. The interpretation of an utterance mostly consists in the detection of application-specific entities. In an application like the France Telecom 3000 Voice Agency service this detection is performed by hand-crafted specific knowledge.

The AT&T *HMIHY* corpus was the first large dialogue corpus, obtained from a deployed system, that has the above mentioned characteristics. A service like the France Telecom 3000 Voice Agency service has been developed by a user interface development lab. This new kind of systems and dialogue corpora offer new opportunities for academic research that can be summarized as follows:

- How can academic research take profit of the system logs of deployed SDS in order to build the *next generation* of SDS, although the dialogues collected have a dialogue flow constrained by the *previous SDS generation*?

- On the other side, what immediate benefits can academic research offer for the improvement of deployed system, while waiting for the *next SDS generation*?

This paper addresses these aspects in the framework of the deployed FT 3000 Voice Agency service. Section 3 presents how the ASR process can be modified in order to detect and reject Out-Of-Domain utterances, leading to an improvement in the understanding performance without modifying the system. Section 4 shows how the FT 3000 corpus can be used in order to build stochastic models that are the basis of a new Spoken Language Understanding strategy, even if the current SLU system used in the FT 3000 service is not stochastic. Section 5 presents experimental results obtained on this corpus justifying the need of a tighter integration between the ASR and the SLU models.

2 Description of the France Telecom 3000 Voice Agency corpus

The France Telecom 3000 (*FT3000*) Voice Agency service, the first deployed vocal service at France Telecom exploiting natural language technologies, has been made available to the general public in October 2005. *FT3000* service enables customers to obtain information and purchase almost 30 different services and access the management of their services. The continuous speech recognition system relies on a bigram language model. The interpretation is achieved through the *Verbateam* two-steps semantic analyzer. *Verbateam* includes a set of rules to convert the sequence of words hypothesized by the speech recognition engine into a sequence of concepts and an inference process that outputs an interpretation label from a sequence of concepts.

2.1 Specificities of interactions

Given the main functionalities of the application, two types of dialogues can be distinguished. Some users call FT 3000 to activate some services they have already purchased. For such demands, users are rerouted toward specific vocal services that are dedicated to those particular tasks. In that case, the *FT3000* service can be seen as a unique automatic frontal desk that efficiently redirects users. For such dialogues the collected corpora only contain the interaction prior to rerouting. It can be observed in that case that users are rather familiar to the system and are most of the time regular users. Hence, they are more likely to use short utterances, sometimes just keywords and the interaction is fast (between one or two dialogue turns in order to be redirected to the demanded specific service).

Such dialogues will be referred as *transit* dialogues and represent 80% of the calls to the *FT3000*

service. As for the 20% other dialogues, referred to as *other*, the whole interaction is proceeded within the *FT3000* application. They concern users that are more generally asking for information about a given service or users that are willing to purchase a new service. For these dialogues, the average utterance length is higher, as well as the average number of dialogue turns.

	other	**transit**
# dialogues	350	467
# utterances	1288	717
# words	4141	1454
av. dialogue length	3.7	1.5
av. utterance length	3.2	2.0
OOV rate (%)	3.6	1.9
disfluency rate (%)	2.8	2.1

Table 1: Statistics on the *transit* and *other* dialogues

As can be observed in table 1 the fact that users are less familiar with the application in the *other* dialogues implies higher OOV rate and disfluency rate[1]. An important issue when designing ASR and SLU models for such applications that are dedicated to the general public is to be able to handle both naive users and familiar users. Models have to be robust enough for new users to accept the service and in the meantime they have to be efficient enough for familiar users to keep on using it. This is the reason why experimental results will be detailed on the two corpora described in this section.

2.2 User behavior and OOD utterances

When dealing with real users corpora, one has to take into account the occurrence of Out-Of-Domain (OOD) utterances. Users that are familiar with a service are likely to be efficient and to strictly answer the system's prompts. New users can have more diverse reactions and typically make more comments about the system. By comments we refer to such cases when a user can either be surprised *what am I supposed to say now?*, irritated *I've already said that* or even insulting the system. A critical aspect for *other* dialogues is the higher rate of comments uttered by users. For the *transit* dialogues this phenomenon is much less frequent because users are familiar to the system and they know how to be efficient and how to reach their goal. As shown in table 2, 14.3% of the *other* dialogues contain at least one OOD comment, representing an overall 10.6% of utterances in these dialogues.

	other	**transit**
# dialogues	350	467
# utterances	1288	717
# OOD comments	137	24
OOD rate (%)	10.6	3.3
dialogues with OOD (%)	14.3	3.6

Table 2: Occurrence of Out-Of-Domain comments on the *transit* and *other* dialogues

Some utterances are just comments and some contain both useful information and comments. In the next section, we propose to detect these OOD sequences and to take this phenomenon into account in the global SLU strategy.

3 Handling Out-Of-Domain utterances

The general purpose of the proposed strategy is to detect OOD utterances in a first step, before entering the Spoken Language Understanding (SLU) module. Indeed standard Language Models (LMs) applied to OOD utterances are likely to generate erroneous speech recognition outputs and more generally highly noisy word lattices from which it might not be relevant and probably harmful to apply SLU modules.

Furthermore, when designing a general interaction model which aims at predicting dialogue states as proposed in this paper, OOD utterances are as harmful for state prediction as can be an out-of-vocabulary word for the prediction of the next word with an n-gram LM.

This is why we propose a new composite LM that integrates two sub-LMs: one LM for transcribing in-domain phrases, and one LM for detecting and deleting OOD phrases. Finally the different SLU strategies proposed in this paper are applied only to the portions of signal labeled as in-domain utterances.

[1] by disfluency we consider here false starts and filled pauses

3.1 Composite Language Model for decoding spontaneous speech

As a starting point, the comments have been manually annotated in the training data in order to easily separate OOD comment segments from in-domain ones. A specific bigram language model is trained for these comment segments. The comment LM was designed from a 765 words lexicon and trained on 1712 comment sequences.

This comment LM, called LM^{OOD} has been integrated in the general bigram LM^{G}. Comment sequences have been parsed in the training corpus and replaced by a *_OOD_* tag. This tag is added to the general LM vocabulary and bigram probabilities $P(\_OOD\_|w)$ and $P(w|\_OOD\_)$ are trained along with other bigram probabilities (following the principle of *a priori* word classes). During the decoding process, the general bigram LM probabilities and the LM^{OOD} bigram probabilities are combined.

3.2 Decision strategy

Given this composite LM, a decision strategy is applied to select those utterances for which the word lattice will be processed by the SLU component. This decision is made upon the one-best speech recognition hypotheses and can be described as follows:

1. If the one-best ASR output is a single *_OOD_* tag, the utterance is simply rejected.

2. Else, if the one-best ASR output contains an *_OOD_* tag along with other words, those words are processed directly by the SLU component, following the argument that the word lattice for this utterance is likely to contain noisy information.

3. Else (i.e. no *_OOD_* tag in the one-best ASR output), the word-lattice is transmitted to further SLU components.

It will be shown in the experimental section that this pre-filtering step, in order to decide whether a word lattice is worth being processed by the higher-level SLU components, is an efficient way of preventing concepts and interpretation hypothesis to be decoded from an uninformative utterance.

3.3 Experimental setup and evaluation

The models presented are trained on a corpus collected thanks to the *FT3000* service. It contains real dialogues from the deployed service. The results presented are obtained on the test corpus described in section 2.

The results were evaluated according to 3 criteria: the Word Error Rate (WER), the Concept Error Rate (CER) and the Interpretation Error Rate (IER). The CER is related to the correct translation of an utterance into a string of basic concepts. The IER is related to the global interpretation of an utterance in the context of the dialogue service considered. Therefore this last measure is the most significant one as it is directly linked to the performance of the dialogue system.

IER	**all**	**other**	**transit**
size	2005	717	1288
LM^{G}	16.5	22.3	13.0
LM$^{G\ +\ OOD}$	15.0	18.6	12.8

Table 3: Interpretation error rate according to the Language Model

Table 3 presents the IER results obtained with the strategy **strat1** with 2 different LMs for obtaining $\hat{W}$: LM^{G} which is the general word bigram model; and LM$^{G\ +\ OOD}$ which is the LM with the OOD comment model. As one can see, a very significant improvement, 3.7% absolute, is achieved on the *other* dialogues, which are the ones containing most of the comments. For the *transit* dialogues a small improvement (0.2%) is also obtained.

4 Building stochastic SLU strategies

4.1 The FT3000 SLU module

The SLU component of the *FT3000* service considered in this study contains two stages:

1. the first one translates a string of words $W = w_1, \ldots, w_n$ into a string of elementary concepts $C = c_1, \ldots, c_l$ by means of hand-written regular grammars;

2. the second stage is made of a set of about 1600 inference rules that take as input a string of concepts C and output a global interpretation γ of

a message. These rules are ordered and the first match obtained by processing the concept string is kept as the output interpretation.

These message interpretations are expressed by an attribute/value pair representing a function in the vocal service.

The models used in these two stages are manually defined by the service designers and are not stochastic. We are going now to present how we can use a corpus obtained with such models in order to define an SLU strategy based on stochastic processes.

4.2 Semantic knowledge representation

The actual *FT3000* system includes semantic knowledge represented by hand-written rules. These rules can also be expressed in a logic form. For this reason, some basic concepts are now described with the purpose of showing how logic knowledge has been integrated in a first probabilistic model and how it can be used in a future version in which optimal policies can be applied.

The semantic knowledge of an application is a *knowledge base* (KB) containing a set of logic formulas. Formulas return truth and are constructed using constants which represent objects and may be typed, *variables*, *functions* which are mappings from tuples of objects to objects and *predicates* which represent relations among objects. An *interpretation* specifies which objects, functions and relations in the domain are represented by which symbol. Basic *inference problem* is to determine whether $KB \models F$ which means that KB entails a formula F.

In SLU, interpretations are carried on by binding variables and instantiating objects based on ASR results and inferences performed in the KB. Hypotheses about functions and instantiated objects are written into a Short Term Memory (STM).

A user goal is represented by a conjunction of predicates. As dialogue progresses, some predicates are grounded by the detection of predicate tags, property tags and values. Such a detection is made by the interpretation component. Other predicates are grounded as a result of inference. A user goal G is asserted when all the atoms of its conjunction are grounded and asserted true.

Grouping the predicates whose conjunction is the premise for asserting a goal G_i is a process that goes through a sequence of states: $S_1(G_i), S_2(G_i), \ldots$

Let Γ_k^i be the content of the STM used for asserting the predicates grounded at the *k-th* turn of a dialogue. These predicates are part of the premise for asserting the *i-th* goal.

Let G_i be an instance of the *i-th* goal asserted after grounding all the predicates in the premise.

Γ_k^i can be represented by a composition from a partial hypothesis Γ_{k-1}^i available at turn $k-1$, the machine action a_{k-1} performed at turn $k-1$ and the semantic interpretation γ_k^i i.e.:

$$\Gamma_k^i = \chi\left(\gamma_k^i, a_{k-1}, \Gamma_{k-1}^i\right)$$

$S_k(G_i)$ is an information state that can lead to a user's goal G_i and Γ_k^i is part of the premise for asserting G_i at turn k.

State probability can be written as follows:

$$P\left(S_k(G_i)|Y_k\right) = P\left(G_i|\Gamma_k^i\right) P\left(\Gamma_k^i|Y_k\right) \qquad (1)$$

where $P\left(G_i|\Gamma_k^i\right)$ is the probability that G_i is the type of goal that corresponds to the user interaction given the grounding predicates in Γ_k^i. Y_k is the acoustic features of the user's utterance at turn k.

Probabilities of states can be used to define a belief of the dialogue system.

A first model allowing multiple dialog state sequence hypothesis is proposed in (Damnati et al., 2007). In this model each dialog state correspond to a system state in the dialog automaton. In order to deal with flexible dialog strategies and following previous work (Williams and Young, 2007), a new model based on a Partially Observable Markov Decision Process (POMDP) is currently studied.

If no dialog history is taken into account, $P\left(\Gamma_k^i|Y\right)$ comes down to $P\left(\gamma_k^i|Y\right)$, γ_k^i being a semantic attribute/value pair produced by the Verbateam interpretation rules.

The integration of this semantic decoding process in the ASR process is presented in the next section.

5 Optimizing the ASR and SLU processes

With the stochastic models proposed in section 4, different strategies can be built and optimized. We are interested here in the integration of the ASR and SLU processes. As already shown by previous studies (Wang et al., 2005), the traditional sequential approach that first looks for the best sequence of words

$\hat{W}$ before looking for the best interpretation $\hat{\gamma}$ of an utterance is sub-optimal. Performing SLU on a word lattice output by the ASR module is an efficient way of integrating the search for the best sequence of words and the best interpretation. However there are real-time issues in processing word lattices in SDS, and therefore they are mainly used in research systems rather than deployed systems.

In section 3 a strategy is proposed for selecting the utterances for which a word lattice is going to be produced. We are going now to evaluate the gain in performance that can be obtained thanks to an integrated approach on these selected utterances.

5.1 Sequential *vs.* integrated strategies

Two strategies are going to be evaluated. The first one (*strat1*) is fully sequential: the best sequence of word $\hat{W}$ is first obtained with

$$\hat{W} = \underset{W}{argmax} P(W|Y)$$

Then the best sequence of concepts $\hat{C}$ is obtained with

$$\hat{C} = \underset{C}{argmax} P(C|\hat{W})$$

Finally the interpretation rules are applied to $\hat{C}$ in order to obtain the best interpretation $\hat{\gamma}$.

The second strategy (*strat2*) is fully integrated: $\hat{\gamma}$ is obtained by searching at the same time for $\hat{W}$ and $\hat{C}$ and $\hat{\gamma}$. In this case we have:

$$\hat{\gamma} = \underset{W,C,\gamma}{argmax} P(\gamma|C)P(C|W)P(W|Y)$$

The stochastic models proposed are implemented with a Finite State Machine (FSM) paradigm thanks to the AT&T FSM toolkit (Mohri et al., 2002).

Following the approach described in (Raymond et al., 2006), the SLU first stage is implemented by means of a word-to-concept transducer that translates a word lattice into a concept lattice. This concept lattice is rescored with a Language Model on the concepts (also encoded as FSMs with the AT&T GRM toolkit (Allauzen et al., 2003)).

The rule database of the SLU second stage is encoded as a transducer that takes as input concepts and output semantic interpretations γ. By applying this transducer to an FSM representing a concept lattice, we directly obtain a lattice of interpretations.

The SLU process is therefore made of the composition of the ASR word lattice, two transducers (word-to-concepts and concept-to-interpretations) and an FSM representing a Language Model on the concepts. The concept LM is trained on the *FT3000* corpus.

This strategy push forward the approach developped at AT&T in the *How May I Help You?* (Gorin et al., 1997) project by using richer semantic models than call-types and named-entities models. More precisely, the 1600 Verbateam interpretation rules used in this study constitute a rich knowledge base. By integrating them into the search, thanks to the FSM paradigm, we can jointly optimize the search for the best sequence of words, basic concepts, and full semantic interpretations.

For the strategy *strat1* only the best path is kept in the FSM corresponding to the word lattice, simulating a sequential approach. For *strat2* the best interpretation $\hat{\gamma}$ is obtained on the whole concept lattice.

error	*WER*	*CER*	*IER*
strat1	40.1	24.4	15.0
strat2	38.2	22.5	14.5

Table 4: Word Error Rate (WER), Concept Error Rate (CER) and Interpretation Error Rate (IER) according to the SLU strategy

The comparison among the two strategies is given in table 4. As we can see a small improvement is obtained for the interpretation error rate (IER) with the integrated strategy (*strat2*). This gain is small; however it is interesting to look at the Oracle IER that can be obtained on an n-best list of interpretations produced by each strategy (the Oracle IER being the lowest IER that can be obtained on an n-best list of hypotheses with a perfect Oracle decision process). This comparison is given in Figure 1. As one can see a much lower Oracle IER can be achieved with *strat2*. For example, with an n-best list of 5 interpretations, the lowest IER is 7.4 for *strat1* and only 4.8 for *strat2*. This is very interesting for dialogue systems as the Dialog Manager can use dialogue context information in order to filter such n-best lists.

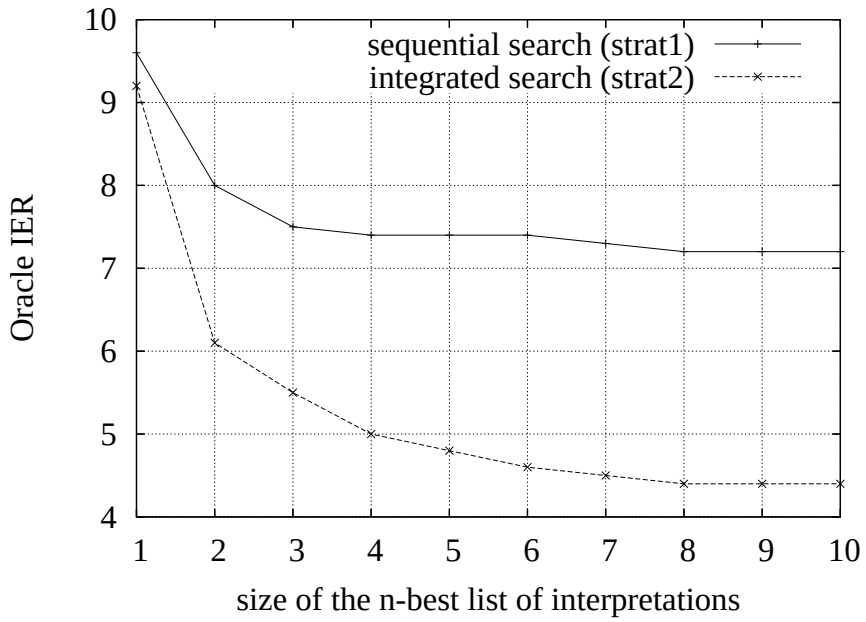

Figure 1: Oracle IER according to an n-best list of interpretations for strategies *strat1* and *strat2*

5.2 Optimizing WER, CER and IER

Table 4 also indicates that the improvements obtained on the WER and CER dimensions don't always lead to similar improvements in IER. This is due to the fact that the improvements in WER and CER are mostly due to a significant reduction in the insertion rates of words and concepts. Because the same weight is usually given to all kinds of errors (insertions, substitutions and deletions), a decrease in the overall error rate can be misleading as interpretation strategies can deal more easily with insertions than deletions or substitutions. Therefore the reduction of the overall WER and CER measures is not a reliable indicator of an increase of performance of the whole SLU module.

level	*1-best*	*Oracle hyp.*
WER	33.7	20.0
CER	21.2	9.7
IER	13.0	4.4

Table 5: Error rates on words, concepts and interpretations for the 1-best hypothesis and for the Oracle hypothesis of each level

These results have already been shown for WER by previous studies like (Riccardi and Gorin, 1998) or more recently (Wang et al., 2003). They are illustrated by Table 5 and Table 6. The figures shown in these tables were computed on the subset of utterances that were passed to the SLU component. Utterances for which an OOD has been detected are discarded. In Table 5 are displayed the error rates obtained on words, concepts and interpretations both on the 1-best hypothesis and on the Oracle hypothesis (the one with the lowest error rate in the lattice). These Oracle error rates were obtained by looking for the best hypothesis in the lattice obtained at the corresponding level (e.g. looking for the best sequence of concepts in the concept lattice). As for Table 6, the mentioned IER are the one obtained when applying SLU to the Oracles hypotheses computed for each level. As one can see the lowest IER (4.4) is not obtained on the hypotheses with the lowest WER (9.8) or CER (7.5).

	IER
from word Oracle	9.8
from concept Oracle	7.5
interpretation Oracle	4.4

Table 6: IER obtained on Oracle hypotheses computed at different levels.

6 Conclusion

This paper presents a study on the *FT3000* corpus collected from real users on a deployed general public application. Two problematics are addressed: How can such a corpus be helpful to carry on research on advanced SLU methods eventhough it has been collected from a more simple rule-based dialogue system? How can academic research translate into short-term improvements for deployed services? This paper proposes a strategy for integrating advanced SLU components in deployed services. This strategy consists in selecting the utterances for which the advanced SLU components are going to be applied. Section 3 presents such a strategy that consists in filtering Out-Of-Domain utterances during the ASR first pass, leading to significant improvement in the understanding performance.

For the SLU process applied to in-domain utterances, an integrated approach is proposed that looks simultaneously for the best sequence of words, concepts and interpretations from the ASR word lattices. Experiments presented in section 5 on real data show the advantage of the integrated approach towards the sequential approach. Finally, section 4 proposes a unified framework that enables to define a dialogue state prediction model that can be applied and trained on a corpus collected through an already deployed service.

References

Cyril Allauzen, Mehryar Mohri, and Brian Roark. 2003. Generalized algorithms for constructing statistical language models. In *41st Annual Meeting of the Association for Computational Linguistics (ACL'03), Sapporo, Japan.*

Helene Bonneau-Maynard, Sophie Rosset, Christelle Ayache, Anne Kuhn, and Djamel Mostefa. 2005. Semantic annotation of the french media dialog corpus. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Lisboa, Portugal.

Geraldine Damnati, Frederic Bechet, and Renato De Mori. 2007. Spoken Language Understanding strategies on the France Telecom 3000 voice agency corpus. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, USA.

A. L. Gorin, G. Riccardi, and J.H. Wright. 1997. How May I Help You ? In *Speech Communication*, volume 23, pages 113–127.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 96–101, Hidden Valley, Pennsylvania.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer, Speech and Language*, 16(1):69–88.

Christian Raymond, Frederic Bechet, Renato De Mori, and Geraldine Damnati. 2006. On the use of finite state transducers for semantic interpretation. *Speech Communication*, 48,3-4:288–304.

Giuseppe Riccardi and Allen L. Gorin. 1998. Language models for speech recognition and understanding. In *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, Sidney, Australia.

Ye-Yi Wang, A. Acero, and C. Chelba. 2003. Is word error rate a good indicator for spoken language understanding accuracy? In *Automatic Speech Recognition and Understanding workshop - ASRU'03*, St. Thomas, US-Virgin Islands.

Ye-Yi Wang, Li Deng, and Alex Acero. 2005. Spoken language understanding. In *Signal Processing Magazine, IEEE*, volume 22, pages 16–31.

Jason D. Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer, Speech and Language*, 21:393–422.