

SPOKEN LANGUAGE UNDERSTANDING STRATEGIES ON THE FRANCE TELECOM 3000 VOICE AGENCY CORPUS

Géraldine Damnati¹, Frédéric Béchet², Renato De Mori²

¹ France Télécom R&D - TECH/SSTP/RVA 2 av. Pierre Marzin 22307 Lannion Cedex 07, France

² LIA - University of Avignon, BP1228 84911 Avignon cedex 09 France

{frederic.bechet,renato.demori}@univ-avignon.fr

geraldine.damnati@orange-ft.com

ABSTRACT

Telephone services are now deployed that allow users to react to telephone prompts in spoken natural language. These systems have limited domain semantics and dialogue strategies which are represented by finite state diagrams. Most of these systems adopt a sequential approach where the Automatic Speech Recognition (ASR) process, the Spoken Language Understanding (SLU) process and the Dialogue Management (DM) are separate processes. In the framework of the France Telecom 3000 voice service, we propose in this paper to study several strategies in order to integrate more closely these three processes: ASR, SLU, and DM. By means of a Finite State Machine paradigm encoding the different models used by these three levels we show how the search for the best sequence of dialogue states can be done simultaneously at the word, concept, interpretation and dialogue state levels.

Index Terms— Automatic Speech Recognition, Spoken Language Understanding, Language Models, Spoken Dialogue Systems

1. INTRODUCTION

Telephone services are now deployed that allow users to react to telephone prompts in spoken natural language. These systems have limited domain semantics and dialogue strategies which are represented by finite state diagrams. State transitions have associated a semantic knowledge represented by logical expressions of conceptual entities. Such a representation is manually derived and is adequate because it has been compiled by experts with a deep knowledge of the domain. Real problems in these systems depend on automatic speech recognition (ASR) errors and on the difficulty in modeling relations between concepts and the way people express them. In order to take these sources of imprecision into account, it is proposed in this paper to conceive a dialogue strategy that considers the possibility that the dialogue is not in a single state at a given phase of its evolution. Rather, dialogue can be in different states and a language generator component generates a prompt to the user that attempts to gather useful information not only for the progress of the dialogue towards a final state but also to reduce the entropy of the information about the actual dialogue state.

In this framework, state transitions are labeled by the fact that the results of ASR causes certain premises to be true and an inference process leads to the truth of derived assertions. As the inference process is guided by inference rules, Finite State Machines (FSM)

are derived from them and plugged into the dialogue Stochastic Finite State Machine (SFSM) in such a way that probabilities of dialogue states can be obtained from word lattice probabilities using operations on automata.

When dealing with real users corpora, one has to be able to handle 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. We propose in this paper to detect such OOD utterances in a first step, before entering into the Spoken Language Understanding (SLU) module. Indeed standard Language Models (LMs) applied to OOD utterances are likely to produce very noisy word lattices from which it might not be relevant to apply SLU modules.

Furthermore, when designing a general interaction model such as the transition state model 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 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.

The paper is organized as follows. In Section 2 the vocal service on which this study has been made is described. Sections 3 and 4 outlines the ASR and SLU decoding processes leading to the computation of state probabilities. In Section 5 details of the interpretation knowledge are provided and experimental results are given in Section 6.

2. DESCRIPTION OF THE FRANCE TELECOM 3000 VOICE AGENCY CORPUS

The **3000** service, the first deployed vocal service at France Telecom exploiting natural language technologies, has been made available to the public in October 2005. **3000** is France Telecom's voice agency that enables customers to obtain information and purchase almost 30 different services, check their consumption, pay their bills and access the management of their services such as call forwarding or voice messaging. 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. Given the main functionalities of the application, two

types of dialogues can be distinguished. Some users dial 3000 to activate some services they have already purchased, such as consumption checking or line transfer. For such demands, users are rerouted towards specific vocal services that are dedicated to those particular tasks. In that case, the **3000** service can be seen as a frontal desk that efficiently redirects users, eventually making use of user profile information. Those 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 two or three 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 3000 service. As for the 20% other dialogues, referred to as *other*, the whole interaction is proceeded within the **3000** 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. Users are less familiar with the application and the disfluency rates as well as the OOV rate are higher.

Another critical aspect for this second type of dialogues is the higher rate of comments uttered by users. By comments we mean utterances that are Out-Of-Domain (OOD). User can either be surprised *what am I supposed to say now?*, irritated *I've already said that* or even insulting the system. For the *transit* dialogues this phenomenon is not frequent because users are familiar to the system and they know how to be efficient and how to reach their goal, but for the *other* dialogues, 10% of the utterances contain such OOD comments. Some utterances are just comments and some contain both useful information and comments. For this purpose we propose in this paper a new strategy that consists in first detecting the comments from the in-domain information, thanks to a composite Language Model, following the one proposed in [1]. The motivation is that such utterances are likely to generate erroneous speech recognition outputs and more generally highly noisy word lattices. It is therefore useless and probably harmful to apply higher level speech understanding techniques to such word lattices.

3. DOMAIN-SPECIFIC 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. Another type of OOD utterances can also occur when users are talking to somebody else during the interaction (e.g. *can you close the door please?*). Those segments are too diverse and are not explicitly modeled here. On the other hand comments to the system tend to have a sufficient redundancy. 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.

For example, if P^G is the probabilities of the general LM and P^{OOD} the probabilities of LM^{OOD} , then the probability of the word string w_1, w_2, w_3, w_4 where the sequence w_2, w_3 is an OOD comment is:

$$P^{G+OOD}(w_1, w_2, w_3, w_4) = P^G(w_1|start) \times P^G(_{OOD_}|w_1) \times P^{OOD}(w_2|start) \times P^{OOD}(w_3|w_2) \times P^{OOD}(end|w_3) \times P^G(w_4|_{OOD_})$$

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.

4. PROBABILITY COMPUTATION OF A SEQUENCE OF STATES IN A FINITE STATE DIALOGUE MODEL

Let us consider a finite state dialogue models (FSDM) in which a transition between a state S_i and a state S_j is labeled by a first order logic expression of concept hypotheses generated by a Spoken Language Understanding system (SLU) which contains an ASR component. A dialogue prompt is generated by the dialogue strategy when the dialogue system is in a state. The user reacts to a dialogue prompt with an utterance which is interpreted by the SUS that generates conceptual interpretations.

As the dialogue progresses, the dialogue strategy visits a sequence of states. Let $S = \{S_0, S_1, \dots, S_k\}$ be such a sequence. State S_k is reached in a dialogue turn in which a sequence of acoustic signal features Y_k have been interpreted leading to a conceptual interpretation Γ_k . This interpretation is obtained by applying a set of logical rules to the conceptual interpretation of Y_k (i.e. the string of basic conceptual entities output by the SLU component). These rules contain predicates, variables and logical operators.

Let $Y = \{Y_1, Y_2, \dots, Y_k\}$ be the sequence of utterance acoustic descriptors and $\Gamma = \{\Gamma_1, \Gamma_2, \dots, \Gamma_k\}$ be the sequence of utterance interpretations. We are interested in computing the probability $P(S|Y)$ and to use it in providing the dialogue strategy with multiple state sequence hypotheses.

A strategy that takes into account multiple state sequence hypotheses has recently been proposed by [2], based on a Partially Observable Markov Decision Process (POMDP). The model proposed here is simpler as all the states are defined by the finite state dialogue model built for the deployed service.

The computation of our model is performed recursively as follows:

$$P(S|Y) = \sum_{\Gamma} P(S\Gamma|Y) = \sum_{\Gamma} P(S_k\Gamma_k|H_kY)P(H_k|Y) \quad (1)$$

with $H_k = \{S_{1,k-1}, \Gamma_{1,k-1}\}$ and

$$\begin{aligned} P(S_k \Gamma_k | H_k Y) &= P(S_k | \Gamma_k H_k Y) P(\Gamma_k | H_k Y) \\ &\approx P(S_k | \Gamma_k H_k) P(\Gamma_k | Y_k) \end{aligned} \quad (2)$$

If no history context is taken into account, then $P(S_k | \Gamma_k H_k)$ is set to $P(S_k | \Gamma_k)$. When a training corpus is available, the history can be approximated by an n-gram model on the dialogue states. In the interpretation module proposed in this paper we use a bigram model, therefore we have:

$$P(S_k | \Gamma_k H_k) \approx P(S_k | \Gamma_k S_{k-1}) \quad (3)$$

Computation of $P(\Gamma_k | Y_k)$ is performed by the following process: the interpretation Γ_k is obtained by applying deterministic logical rules to the string of basic concepts C_k obtained from Y_k . More precisely this string concept C_k is obtained from the word string W_k recognized in Y_k . Therefore we have:

$$\begin{aligned} P(\Gamma_k | Y_k) &= \sum_{C_k, W_k} P(\Gamma_k C_k W_k | Y_k) \\ &\approx \sum_{C_k, W_k} P(\Gamma_k | C_k) P(C_k | W_k) P(W_k | Y_k) \end{aligned}$$

As the interpretation rules are not stochastic, $P(\Gamma_k | C_k)$ is either equal to 1 if the rule leading to Γ matches the string of concepts C_k and 0 otherwise. $P(C_k | W_k)$ is given by a concept tagger, estimating the best sequence of concept thanks to an HMM-based tagger as presented in [3]. $P(W_k | Y_k)$ is given by the ASR models (acoustic+LM).

With this framework several interpretation strategies can be built. We compare in this study 3 different strategies:

1. This first strategy is purely sequential and correspond to most of the SLU systems: the best sequence of word \hat{W} is first obtained thanks to $\hat{W} = \underset{W}{\operatorname{argmax}} P(W | Y)$. Then the best sequence of concepts \hat{C} is obtained with $\hat{C} = \underset{C}{\operatorname{argmax}} P(C | \hat{W})$.

The set of interpretation rules is applied to \hat{C} in order to obtain Γ . No dialogue history is taken into account, therefore equation 2 becomes:

$$P(S_k \Gamma_k | H_k Y) \approx P(S_k | \Gamma_k) P(\Gamma_k | \hat{C}_k) P(\hat{C}_k | \hat{W}_k) P(\hat{W}_k | Y_k)$$

This strategy is called *strat1* in the experiments section.

2. The second strategy looks at the same time for the best sequence of words and concepts. Again no history is taken into account leading to:

$$P(S_k \Gamma_k | H_k Y) \approx P(S_k | \Gamma_k) \times \max_{W_k, C_k} P(\Gamma_k | C_k) P(C_k | W_k) P(W_k | Y_k)$$

This strategy is called *strat2* in the experiments section.

3. The last strategy integrates the dialogue history, therefore strategy *strat3* corresponds to:

$$P(S_k \Gamma_k | H_k Y) \approx P(S_k | \Gamma_k S_{k-1}) \times \max_{W_k, C_k} P(\Gamma_k | C_k) P(C_k | W_k) P(W_k | Y_k)$$

At each dialogue turn k the state sequence $S = \{S_0, S_1, \dots, S_k\}$ is estimated thanks to $P(S | Y)$. The n-best dialogue states are sent to the Dialogue Manager.

5. SPOKEN LANGUAGE UNDERSTANDING COMPONENT

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

1. the first one translates a string of words into a string of elementary concepts;
2. the second stage is made of a set of about 1600 inference rules that take as input a string of concepts and output a global interpretation 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 presented in the previous section are implemented with a Finite State Machine (FSM) paradigm thanks to the AT&T FSM toolkit [4]. Following previous work described in [5], the first stage is implemented by means of a word-to-concept transducer that translates a word lattice into a concept lattice. The rule database of the second stage is also encoded as a transducer that takes as input concepts and output rule identification number (corresponding to the position of the rule in the database). The SLU process is therefore made of the composition of the ASR word lattice and the two transducers: word-to-concepts and concept-to-interpretations. For the strategies *strat2* and *strat3*, two Language Models (also encoded as FSMs with the AT&T GRM toolkit [6]) are added to the composition operation: a LM on concepts for *strat2* and a LM on dialogue states for *strat3*.

6. EXPERIMENTS

The models presented in section 5 are trained on a corpus collected thanks to the France Telecom 3000 Voice Agency service. It contains real dialogues from the deployed service. The concept tagger is trained on a set of 44K utterances manually transcribed and conceptually labeled. The LM on the dialogue states is trained on a 7.4K dialogue corpus. The results presented are obtained on a test corpus made of 816 dialogues and 1953 utterances (or dialogue turn). This corpus contains 1219 utterances corresponding to dialogues labeled as *transit* as presented in section 2, and 734 utterances for the *other* dialogues.

The results are given 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 and is computed as the Word Error Rate with the same weight for an insertion, a deletion or a substitution of a concept. The IER is related to the global interpretation of an utterance in the context of the dialogue service considered. There is one interpretation label for each utterance, containing the predicate and the attributes representing the users requests. 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	1953	734	1219
LM ^G	16.5	22.3	13.0
LM ^{G+OOD}	15.0	18.6	12.8

Table 1. Interpretation error rate according to the Language Model

Table 1 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.

corpus	all		
error	WER	CER	IER
strat1	40.1	24.4	15.0
strat2	38.2	22.5	14.5
strat3	38.3	22.5	14.7
corpus	other		
error	WER	CER	IER
strat1	48.8	34.7	18.6
strat2	47.6	34.2	18.9
strat3	47.9	34.4	19.4
corpus	transit		
error	WER	CER	IER
strat1	31.8	18.2	12.8
strat2	29.3	14.2	11.8
strat3	29.1	14.0	11.8

Table 2. Word Error Rate (WER), Concept Error Rate (CER) and Interpretation Error Rate (IER) according to the SLU strategy

The comparison among the different strategies is given in table 2. 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. These results have already been shown for WER by previous studies like [7] or more recently [8].

By considering only the IER measure, table 2 shows that a small gain in performance is achieved with **strat2**, specially for the *transit* dialogues. **strat3** on the other hand doesn't bring any gain. This can be explained by the fact that the dialogues in the FT 3000 Voice Agency corpus are short dialogues, therefore the dialogue history is not a discriminant feature in these experiments. However it is interesting to compare the interpretations obtained by the 3 strategies: by considering the consensus among them, one can see that adding new source of information to the baseline model, as in **strat2** and **strat3**, is an efficient way for detecting problematic dialogues. For example table 3 shows the IER obtained by keeping only the utterances that lead to a consensus in the interpretations obtained with the 3 strategies. The coverage in term of corpus is also displayed. With the 3 strategies, the IER decreases from 15.0% to 12.0% and the utterances kept cover 87.6% of the test corpus.

7. CONCLUSION

This study presents interpretation results obtained with the France Telecom 3000 voice agency corpus. The SLU process proposed is

IER	all		other		transit	
consensus	IER	cover	IER	cover	IER	cover
1	15.0	100%	18.6	100%	12.8	100%
1^2	12.7	88.7%	15.1	86.4%	8.7	92.8%
1^2^3	12.0	87.6%	14.3	84.9%	8.3	92.3%

Table 3. IER according to consensus among the strategies

a unified search process that looks simultaneously for the best sequence of words, concepts, interpretations and dialogue states. After filtering the messages in order to discard out-of-domain phrases, like users comments, the word lattices corresponding to the in-domain utterances are composed with several FSMs representing both semantic structures, like the concept tagger and the interpretation rules, and LMs on concepts and dialogue states.

The results make evident the need for SLU evaluation to cover the complete interpretation of a message rather than the error rate on the basic constituents. They also show that the approach proposed is promising as this SLU strategy is an efficient way to add dialogue context information into the search process of the best interpretation of a message.

8. REFERENCES

- [1] Nathalie Camelin, Geraldine Damnati, Frederic Bechet, and Renato De Mori, "Opinion mining in a telephone survey corpus," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh, PA, USA, 2006, pp. 1041–1044.
- [2] Jason D. Williams and Steve Young, "Partially observable markov decision processes for spoken dialog systems," *Computer, Speech and Language*, vol. 21, pp. 393–422, 2007.
- [3] Christophe Servan, Christian Raymond, Frederic Bechet, and Pascal Nocera, "Conceptual decoding from word lattices: Application to the spoken dialogue corpus MEDIA," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh, PA, USA, 2006, pp. 1614–1617.
- [4] Mehryar Mohri, Fernando Pereira, and Michael Riley, "Weighted finite-state transducers in speech recognition," *Computer, Speech and Language*, vol. 16, no. 1, pp. 69–88, 2002.
- [5] Christian Raymond, Frederic Bechet, Renato De Mori, and Geraldine Damnati, "On the use of finite state transducers for semantic interpretation," *Speech Communication*, vol. 48,3-4, pp. 288–304, 2006.
- [6] Cyril Allauzen, Mehryar Mohri, and Brian Roark, "Generalized algorithms for constructing statistical language models," in *41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*, Sapporo, Japan, 2003.
- [7] Giuseppe Riccardi and Allen L. Gorin, "Language models for speech recognition and understanding," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Sidney, Australia, 1998.
- [8] Ye-Yi Wang, A. Acero, and C. Chelba, "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, 2003.