SPEAKER TURN CHARACTERIZATION FOR SPOKEN DIALOG SYSTEM MONITORING AND ADAPTATION

Geraldine Damnati[1], *Frederic Bechet*[2], *Renato de Mori*[2]

[1]France Telecom R&D - Orange Labs - 2 av. Pierre Marzin 22307 Lannion, France
[2]Université d'Avignon - LIA - 339 ch. des Meinajaries 84911 Avignon, France

geraldine.damnati@orange-ftgroup.com, frederic.bechet@univ-avignon.fr, renato.demori@univ-avignon.fr

ABSTRACT

This paper describes an utterance classification method based on a multiple decoding scheme. We use the Spoken Language Understanding (SLU) strategy proposed within the European project LUNA. The goal of this classification process is to characterize each speaker's turn, in a dialog context, according to different categories relevant from an SLU point of view: out-of-domain messages, requests not covered by the interpretation module, frequent requests,.... These categories are used for two purposes in an off-line mode: system monitoring for detecting changes in users' behaviour and system adaptation by selecting dialogs likely to contain some phenomenon poorly covered by the models for an active learning scheme. All the models and the evaluations are performed on the France Telecom *FT3000* corpus.

Index Terms— speech recognition, spoken language understanding, spoken dialog system, active learning

1. INTRODUCTION

Spoken Dialog Systems (SDS) are now deployed on a large scale for full or partial automation in call-centers such as customer care services [1, 2, 3]. Such applications are usually deployed thanks to an initial training corpus of limited size, acquired by various wizard-of-oz techniques [4]. This initial corpus can be used for two purposes:

- extracting global knowledge on the application targeted such as concept ontologies, manual grammars or interpretation rules;
- estimating distributions on word, concept and interpretation for training language or classification models, if this initial corpus is big enough for extracting reliable distributions.

Once the application is deployed, new examples collected on a daily basis are gathered in order to build a second corpus, containing real interactions from this initial deployed system. This new corpus can be used to update the initial models (knowledge based, or learnt ones). Another use of this collected corpus is to monitor the deployed system in order to collect statistics on its use and estimate its performance without requiring the manual annotation of a significant portion of dialogs [5, 2].

However this kind of collected corpus contains two biases :

- firstly, since the dialogs are collected with an existing system that has its own limitations, the collected corpus will reflect these limitations and therefore will not necessarily be directly useful for building a more advance application;
- secondly, if the service is widely deployed, frequent users are going to adapt themselves to the existing service, learning which formulations lead to an efficient process of their requests. If this is very good from the service efficiency point of view, the corpus collected from them contains little unseen events. Moreover the distributions of word, concept and interpretation extracted from such dialogs will overwhelm the distributions of rare or atypical requests, potentially the most interesting ones from the system improvement point of view.

Therefore it is important to select, from all the dialog traces available, those that are the most likely to improve the system performance. This corresponds to an active learning approach [6] that consists in selecting dialogs with automatic criteria. These selected samples, manually processed, can be used for updating knowledge-based models, for example by increasing the coverage of interpretation rules. They can also be added to the training corpus of statistical processes.

This paper addresses this problem of exploiting collected dialog traces such as log files from deployed Spoken Dialog Systems (SDS) for system monitoring and adaptation. This study is done within the Spoken Language Understanding (SLU) strategy proposed by the European project LUNA. The different interpretation models proposed in LUNA are used to characterize speaker turns, in a dialog context, according to a set of predefined situations that are used for selecting relevant utterances for system diagnosis and active learning strategies.

Section 2 presents the corpus on which this study has been made; section 4 presents the multiple decoding scheme proposed in this paper for characterizing speaker turns and section 5 evaluates this characterization process.

2. THE FT3000 CORPUS

The *FT3000* service is the first Natural Language Application deployed at France Telecom processing non constraint spontaneous speech. It was launched in October 2005. This service allows France Telecom customers to obtain information about their bills, register to over 30 different services or manage their France Telecom account.

The semantic model is made of 400 different concepts (named entities, dialog command, keywords) leading to 2030 possible structured interpretations, represented as semantic frames. The Dialog Manager (DM) implements a finite state approach: at each dialog turn several states can be reached, each of them called a *phase*. There

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.

are 137 phases in the DM. To each phase is attached the set of frames or interpretations that leads to another transition in the dialog automaton. These are the *expected* frames according to the dialog context. Any other frame recognized will be ignored by the system and if no valid interpretations are recognized in a user turn, the message is rejected.

The *FT3000* corpus contains dialog log files between the deployed system and real France Telecom customers. Therefore this corpus contains the whole range of realistic issues often missing in lab corpora: unwanted cut communication, noises, non-cooperative or even angry users, ...

In order to analyse the performance of a SLU system, it is important not only to evaluate its global performance but also to distinguish the different kinds of messages the system is faced to. The first distinction is among the messages that contain or not a valid interpretation according to the global and the local dialog context. If no valid interpretation is found, the message has to be rejected. In the FT3000 corpus, 22% of the messages is in this case and has to be rejected. This shows that improving the performance of an SLU system is also linked to its capacity to reject non valid messages.

These rejected messages are either really *empty* messages (non speech signal) or they might contain some linguistic input that cannot be parsed by the SLU modules. Some of them can be very useful to add to the training corpus in order to augment the coverage of the SLU models. It is therefore very important to characterize each message with some information that can be used in such an active learning process. This characterization is presented in the next section.

3. CHARACTERIZING SPOKEN UTTERANCES

We define several categories of spoken utterances. The set of messages to be rejected is divided in three categories and the set of messages containing a valid interpretation is split in two categories. Therefore we consider five user turn categories in this study:

1. C_1 contains the messages with no linguistic input (noise, silence) ;
2. C_2 contains out-of-domain utterances, which can be comments from users, appreciations, or even utterances related to something out of the focus of the FT3000 corpus with therefore only Out-Of-Vocabulary words ;
3. C_3 contains utterances related to the service but which don't lead to any valid interpretation because either the request is partial or it addresses functionalities not covered by the deployed system ;
4. C_4 contains messages with valid interpretation but related to low-frequency requests in the current dialog context ;
5. C_5 contains the frequent valid interpretations, therefore this category is the most frequent one in the *FT3000* corpus.

The distinction between C_4 and C_5 is based on the frequency of the pairs phase/interpretation found in the training corpus. There are 27667 such pairs and a pair is considered as frequent (category C_5) if it appears more than a hundred times in the corpus. This covers 59% of the occurrences of pairs although it contains only 37 distinct pairs among the 1854 different ones found in the training corpus. This shows that most users of the *FT3000* service use only a limited set of functionalities that are over represented in the training corpus. Monitoring C_5 utterances over time is a way of detecting the arising of a new problem (typically if the proportion of infrequent requests becomes abnormally high). On the other hand, retrieving efficiently C_3 utterances is helpful for improving the application coverage, in an active learning perspective for instance.

Table 1 shows the distributions of the messages among the 5 categories on a corpus collected during a 10 day period with the *FT3000* service. This corpus is a snapshot of the deployed system. As we can see, 22.3% of the messages should be rejected. This illustrates the main different between corpora collected in lab condition, like ATIS, which contain only C_4 and C_5 messages and those collected with deployed services.

4. THE LUNA SPOKEN LANGUAGE UNDERSTANDING STRATEGY

As presented in [3] and illustrated in figure 1, the SLU process developed in LUNA on the *FT3000* corpus is a 3-level process:

1. The first level translates a word lattice into a concept lattice by means of a Finite State Machine (FSM) transducer containing all the local grammars representing the *FT3000* concepts. This transducer is built from the concept definitions obtained with expert-based rules and belongs to the models *M_init*. Then all the paths of this word/concept transducer can be scored thanks to a Hidden Markov Model tagger trained on the corpus collected from the deployed system. This tagger belongs to *M_freq*. The concept distributions estimated on this corpus for each dialog state can be added to the *M_local* models.
2. The second level applies logical rules on the concept strings in order to build structured interpretations. There are about 2600 manual rules for the *FT3000* service, they are also represented as a FSM transducer translating a concept string into a structured interpretation. The rules and the priority weight given to each of them have been developed by the system designers and belong to *M_init*. The global and local interpretation distributions are then estimated on the collected corpus and added to the models *M_freq* and *M_local*.
3. The third level is the interpretation selection process that chooses among all the possible interpretations the one that fits best the current dialog state. This selection is made according to rules defined in the Dialog Manager (*M_init* and *M_local*) as well as corpus-based decision strategies based on classifiers trained on the collected corpus (*M_freq*).

We group all the SLU models into three categories: *M_init* contains all the knowledge-based hand-crafted models produced by the designer of the system, they define the *acceptability* of a given interpretation ; *M_freq* represents all the models trained on a corpus obtained after deploying the system and they define the *plausibility* of an interpretation ; finally the *M_local* models contains all the information linked to a given dialogue state or *phase*.

As presented in the previous section, half of the messages that has to be rejected contains only noise or out-of-domain (and therefore out-of-vocabulary) content. These messages correspond to the categories C_1 and C_2 and represent 10.5% of the messages. These messages are detected during the ASR process by both a rejection model, in charge of detecting out-of-vocabulary words, and a specific language model dedicated to recognize comments expressed by the users. This rejection strategy is described in [3]. These rejection models are not specific to a single dialog application and are considered as part of the *M_init* models.

Only the models *M_init* are mandatory in the LUNA SLU strategy. The others are optional. By using all or only a combination of

Cat.	nb	%	example
C_1	246	5.4%	*noise*
C_2	231	5.1%	*well what am I supposed to say now*
C_3	538	11.8%	*I receive calls all the time and there's nobody I don't know who it is*
C_4	1870	41.0%	*I'm calling in order to have information about uh for having the wake-up service*
C_5	1671	36.7%	*pay my bill*

Table 1. Utterance categories on the corpus *FT3000*

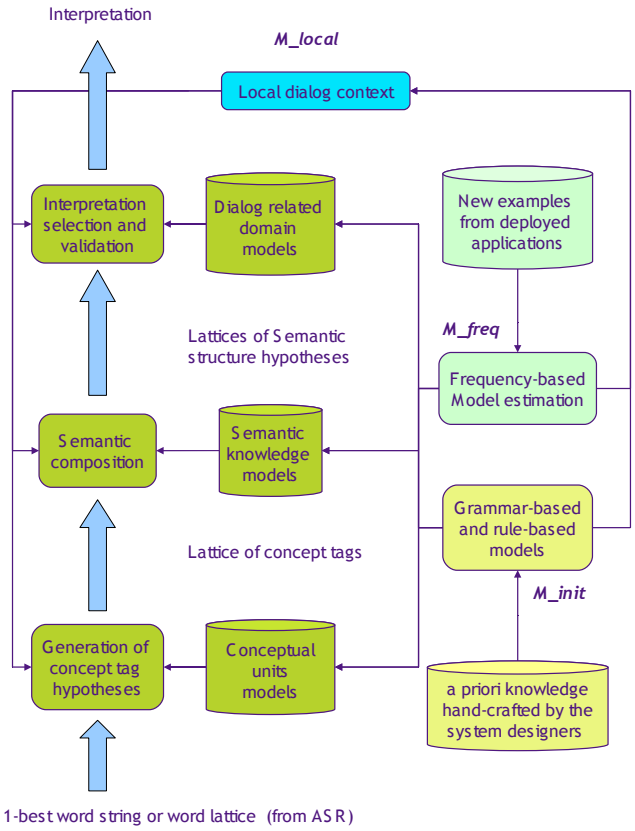

Fig. 1. The LUNA SLU strategy

them we can build different SLU strategies, according to the required system behaviour. For example by using the *M_freq* models we will better recognize the most current requests but with a negative impact on atypical ones. Adding local context can increase the recognition rate of *standard* dialogues but will lead to more difficulties for dialogues already going wrong.

There is no optimal strategy: according to the customer, the request, the audio quality, all these models can be added or removed in order to fit best the current interaction. Let's point out that this decoding scheme is applied only on a limited search space produced by the ASR module, and as all the models implement an FSM approach, performing the different decoding strategies has no impact on the processing time of a message. In a monitoring or active learning perspective, combinations of these systems can be applied offline on large audio databases in order to accurately extract relevant categories of spoken utterances.

In the study presented in this paper we run simultaneously three strategies according to the kind of model involved:

1. S_1=*M_init*
2. S_2=*M_init*+*M_freq*
3. S_3=*M_init*+*M_freq*+*M_local*

Each strategy S_i produces an interpretation hypothesis H_i. By looking at the agreement situations between the hypothesis H_i obtained with our multiple views decoding scheme, we can characterize each dialog turn.Agreement situations can be used in order to focus the selection on messages likely to contain the phenomenon that we want to model with a better accuracy. This process can also estimate in real time the distributions of the messages among the different categories, giving a snapshot of the SDS activities. A noticeable change in these distributions can indicate a change in the users' behaviour that can be linked to a problem arising in the system.

5. EVALUATION

The training corpus used to train the *M_freq* models is made of 42k utterances manually transcribed and automatically annotated thanks to the *FT3000* SLU system. The ASR language model is also trained on this corpus, with an ASR lexicon of 2.2K words. The average Word Error Rate of the 1-best word string produced by the ASR module is 38%. Another corpus containing 4554 user turns collected in May 2007 has been manually transcribed and annotated at the concept and interpretation levels. It is used in this study as a test corpus. This corpus is made of 4.5k utterances containing 7.6 k occurrences of concepts.

5.1. SLU performance of each strategy

The three strategies $S_1 \ldots S_3$ have been applied to the ASR output of each message of this test corpus, producing the three interpretation hypotheses H_1, H_2, H_3. The evaluation of these strategies is given in table 2 on the non-empty messages of the test corpus. An interpretation is considered as correct if the entire frame and all the frame elements (or concepts) are correct. As we can see using the *M_freq* models in S_2 improves the performance. However adding the local context decreases the interpretation accuracy. This can be explained by the fact that local models use the current dialog phase automatically detected. If there is an error in the phase, the local models have a tendency to choose from the lattice of interpretations the hypothesis which better match this erroneous phase.

For the empty messages corresponding to the categories C_1, C_2 and C_3, the rejection strategy based on the models *M_init* gives a precision of 82.7% for a recall of 59.5% on the whole test corpus.

5.2. Message classification

The goal of this study is to characterize each message according to the categories presented in section 3, . In order to estimate the distri-

Hypothesis	H_1	H_2	H_3
% correct interpretation	86.0%	87.8%	86.3%

Table 2. Interpretation evaluation for the three hypotheses H_1, H_2 and H_3 on the non empty messages (3646 messages) of the test corpus.

butions of the message categories of the deployed system, we use a classification process taking as input the different features available in the system log files (like the ASR confidence scores) as well as Boolean features on the agreement of the hypotheses H_1, H_2 and H_3.

We choose the classifier Icsiboost[1], based on the Adaboost algorithm [7]. We performed a 10-fold cross validation process on the *FT3000* test corpus. The results are presented in table 3.

Category C_i	#	precision	recall	F-mes
C_1	246	53.6	45.9	49.5
C_2	357	38.5	23.8	29.4
C_3	410	48.6	38.8	43.2
C_4	1870	85.7	93.5	89.4
C_5	1671	91.6	96.2	93.8
Total	4554	81.5	81.5	81.5

Table 3. Classification results on the 5 categories $C_1 \dots C_5$ on the *FT3000* test corpus

As expected the results are very good for C_5, as this category represents the most frequent requests expressed by the callers, which are well represented in the training corpus of the models. The classification results for the *rejection* messages (C_1, C_2 and C_3) show that it is difficult, when a message has been rejected, to find out at which level this rejection happened. However these results are encouraging as the precision achieved for the most interesting category from an active learning point of view, C_3, is almost 50%. This means that half the messages selected this way effectively contain in-domain speech not yet covered by the SLU or ASR models.

The overall classification rate (81.5%) shows that this method can be an effective way to monitor a deployed Spoken Dialog System by detecting any change in the distributions of the users' messages among the different categories.

5.3. Selecting messages for an active learning process

The goal of this message selection process is to detect the messages that have to be manually transcribed and annotated in order to update and adapt the current SLU modules. Therefore we are interested here in the messages badly recognized by the SLU modules. Using only ASR confidence measures for selecting these messages put together all the problematic messages regardless of their nature and complexity. By using our message classification process as well as different agreement situations between the different hypotheses produced by the strategies S_1, S_2 and S_3, we can achieve an efficient message selection process.

The main target of this selection process is the messages belonging to the category C_3 (in order to augment the coverage of the ASR and SLU models) and the messages belonging to C_4 that are badly recognized. These C_4 messages are interesting because they contain a valid interpretation, according to the current SLU models, but they are rare in the training corpus in the current dialog phase. Therefore collecting more examples of them in order to augment the corpus of the *M_freq* models can help their correct processing.

For collecting the C_3 messages we use directly the message classification process described in the previous section. As already mention this process allows the annotators to retrieve C_3 messages with a precision of 48.6%. This is to compare with the 11.8% that would obtain a random sampling strategy, according to the category distributions given in table 1.

For selecting C_4 messages we use both the classification process (85.7% precision for C_4) and some agreement situations between the hypotheses H_1, H_2 and H_3. If all the hypotheses agree ($H_1 = H_2 = H_3$), the correct interpretation rate reaches 93.9% on the non-empty messages and this corresponds to 78% of the messages. On the contrary, when at least two hypotheses disagree, the correct interpretation rate drops down to 10.9%. Indeed, among the 186 messages of the test corpus belonging to C_4 for which a disagreement situation is observed, 155 of them are wrongly interpreted (83.3%). Therefore this agreement/disagreement rule can be used in order to select for manual annotation the C_4 messages very likely to have been badly processed.

6. CONCLUSION

We have presented in this paper a study on spoken message characterization in a dialog context from the ASR and SLU points of view. We believe this work is a first step in the direction of developing models with a self-diagnosis capacity which can be used either at real-time for adapting decision processes to a given context and monitoring a deployed system, or in an off-line mode for selecting dialogs for manual annotation in an active learning scheme.

7. REFERENCES

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

[2] S. Douglas, D. Agarwal, T. Alonso, RM Bell, M. Gilbert, DF Swayne, and C. Volinsky, "Mining Customer Care Dialogs for Daily News," *IEEE Transactions on Speech, Language and Audio Processing*, vol. 13, no. 5 Part 1, pp. 652–660, 2005.

[3] Geraldine Damnati, Frederic Bechet, and Renato de Mori, "Spoken language understanding strategies on the France Telecom 3000 Voice Agency corpus," in *IEEE ICASSP*, Honolulu, HI, April 2007.

[4] G.D. Fabbrizio, G. Tur, and D. Hakkani-Tür, "Automated Wizard-of-Oz for Spoken Dialogue Systems," in *Ninth European Conference on Speech Communication and Technology*. 2005, ISCA.

[5] A. Abella, J. Wright, and AL Gorin, "Dialog trajectory analysis," in *IEEE ICASSP*, 2004, vol. 1.

[6] D. Hakkani-Tur, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in *IEEE ICASSP*, 2002, vol. 4.

[7] Robert. E. Schapire and Yoram. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, pp. 135–168, 2000.

[1] `http://code.google.com/p/icsiboost/`