

# First implementation of the LUNA Spoken Language Understanding strategy on a telephone service application

Géraldine Damnati<sup>1</sup>, Frédéric Béchet<sup>2</sup>, and Renato De Mori<sup>2</sup>

<sup>1</sup> France Telecom R&D — Orange Labs — 2 av. Pierre Marzin 22307 Lannion, France

<sup>2</sup> University of Avignon — LIA — 339 ch. des Meinajaries 84911 Avignon, France

## Abstract

This paper describes the first results achieved within the LUNA project in coupling the Spoken Language Understanding process with the Automatic Speech Recognition and Dialog Manager processes. This strategy is implemented and evaluated on a France Telecom telephone service application called FT3000.

## 1 Introduction

Since the deployment on a very large scale of the AT&T *How May I Help You?* (HMIHY) service in 2000 (Gorin *et al.*, 1997), Spoken Dialogue Systems (SDS) handling a very large number of calls are now developed from an industrial point of view. The tasks targeted by these deployed SDS are simple, like call-routing or form-filling applications, and the quality of the human-computer interaction is still far from being effective both from the user and service provider side. To improve the effectiveness and user acceptance of automated dialogue systems as well as increasing the complexity of the targeted tasks, one of the main issues is to advance the state of the art of Spoken Language Understanding (SLU) research along two directions:

- Firstly, SLU models have to be tightly coupled with the upstream (Automatic Speech Recognition: ASR) and downstream (Dialog Management: DM) processes.
- Secondly, SLU models have to be part of an adaptive component whose parameters are updated on-line based on the outcome of dialog strategies, a-priori or a-posteriori knowledge.

The goal of the LUNA European project<sup>1</sup> is to develop research along these two directions. This paper describes the first results achieved in coupling the SLU process with the ASR and DM processes. The LUNA strategy is implemented and evaluated on a France Telecom telephone service application called FT3000. Section 2 summarizes the LUNA strategy; section 3 describes the application and the corpus used in this study: the semantic and contextual models are presented in sections 4 and 5 and evaluated in section 6.

---

<sup>1</sup><http://www.ist-luna.eu/>

## 2 The LUNA Spoken Language Understanding strategy

Spoken Language Understanding (SLU) consists in mapping speech signals into conceptual representations. Such complex transduction is traditionally decomposed into two sub-problems. Firstly, spoken utterances are decoded into word hypotheses: word string, set of word strings (n-best) or word lattices. Secondly, such word hypotheses are mapped into conceptual representations.

The conceptual structures used in LUNA for representing a message interpretation are modelled as the composition of elementary semantic constituents. These structures are encoded as Semantic Frames (Baker *et al.*, 1998).

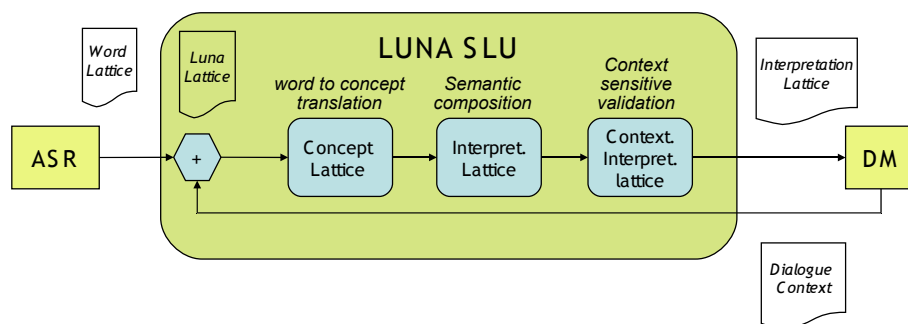


FIGURE 1: The LUNA pipeline: dataflow between the different SLU components.

One of the goal of LUNA is to propose SLU models that are tightly coupled with the upstream (Automatic Speech Recognition: ASR) and downstream (Dialogue Management: DM) processes. The integration of the SLU and ASR processes is realized here by means of the LUNA lattice format that allows a set of alternative hypotheses to be kept at each level of the SLU process. The integration with the downstream (DM) process is done by means of contextual information (local or global) provided by the DM at each SLU processes in order to constrain or score the different conceptual and semantic hypotheses.

Figure 1 presents an overview of the LUNA SLU strategy. A word lattice output by the ASR process is projected into a concept lattice by a translation process. The basic concepts obtained are composed into semantic structures, producing a lattice of interpretations. These interpretations are evaluated and eventually modified by a context-sensitive validation module and the final result is a lattice of contextual interpretations that can be processed by the dialogue manager.

The dataflow between these different stages is presented in figure 1.

Figure 2 gives a description of the different levels of information available in order to implement this LUNA SLU strategy as well as the different models involved.

As we can see, for the three levels of the SLU process, three kinds of information can be used:

- *global* information coming from the *a priori* knowledge provided by the design-

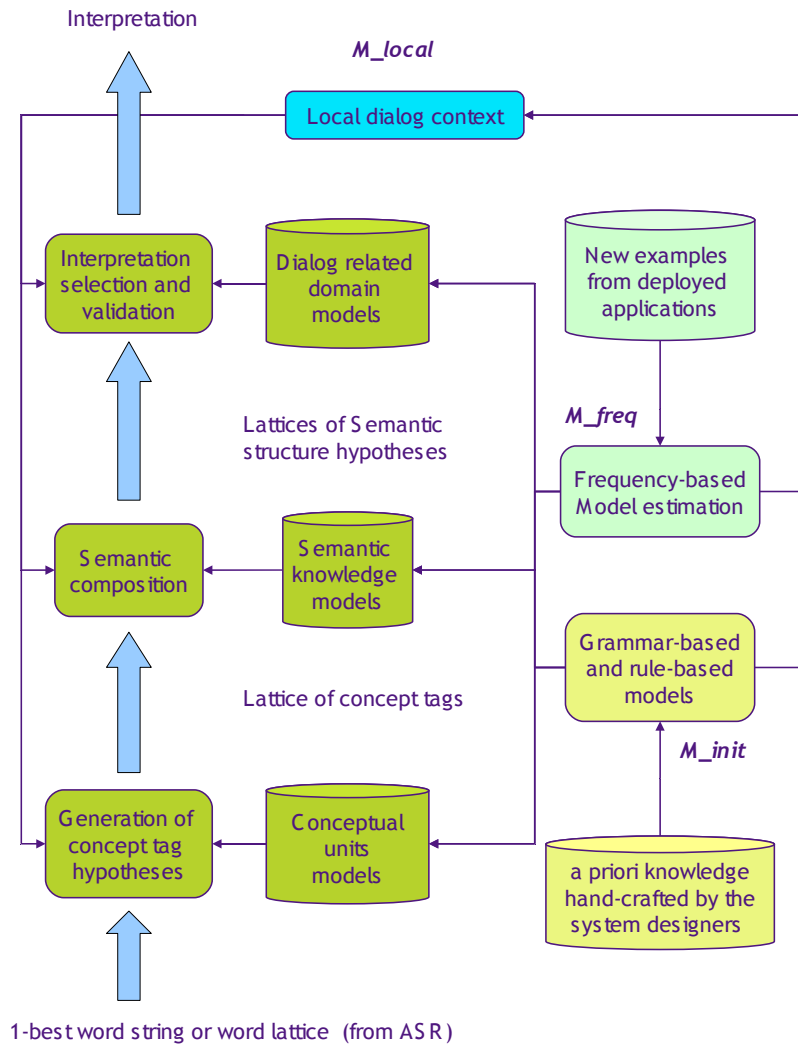


FIGURE 2: The LUNA SLU strategy

ers of the SDS, the models  $M_{init}$  are obtained on the basis of expert-based rules;

- *global* information coming from a corpus collected on the deployed system, represented by statistics on users messages and requests, allowing the system designers to train statistical Language and Classification models ( $M_{freq}$ );
- *local* information, linked to the current dialogue state and provided by the dialogue manager, that can be either knowledge about what is acceptable at

a current dialog state, or statistics about the distributions on words, concepts and interpretations at this state (*M\_local*).

*M\_init* contains all the models obtained during the development and the deployment 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*.

### 3 The France Telecom 3000 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 France Telecom two-steps semantic analyzer. This semantic analyser 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. The FT3000 corpus used in this study is made of about 5000 dialogue turns, manually transcribed and semantically annotated with the first three levels of annotation defined within the LUNA project (Raymond et al., 2007): words, concepts and semantic frames.

### 4 Semantic Model

The semantic model used for the FT3000 corpus contains two levels:

1. the first one translates a string of words  $W = w_1, \dots, w_n$  into a string of elementary concepts  $C = c_1, \dots, c_m$  by means of hand-written regular grammars;
2. the second level is made of a set of about 2600 inference rules that take as input a string of concepts  $C$  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 implementation of this semantic model to the LUNA SLU strategy has been made as follows: both the concept grammars and the inference rules have been represented by Finite State Transducers (FST) thanks to the AT&T FSM library (Mohri et al., 1997). For the concept grammars the FST takes as input words and output concept tags. These concept tags are accepted by the inference rule FST that output message interpretations as well as the rank of the rule used to produce it. The LUNA pipeline is then implemented by a composition operation between 3 different automata: the ASR word lattice represented by a Finite State Machine with word symbols, the concept FST and the inference rule FST. More details on this implementation can be found in Damnati et al. (2007). The local

and global context information provided by the DM is integrated in this process by adding constraints on the generation of concept tags and interpretations and also in the weight calculation of the different paths in the FST obtained after the composition process. This contextual information is described in the next section.

## 5 A multiple view decoding scheme

As presented in Damnati *et al.* (2007) and illustrated in figure 2, 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 an Hidden Markov Model tagger, as presented in Damnati *et al.* (2007), 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*).

All the models used in this study are represented with FSMs by means of the AT&T *FSM Library* for the transducers and *GRM Library* for the language models and the HMM concept tagger. As presented in figure 2 we group these 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*.

Only the models *M\_init* are mandatory in the LUNA SLU strategy. The others are optional. By using all or only a combination of them we can build different SLU strategies, according to the required system behaviour. For example, if we process with *M\_init* only the best word string produced by the ASR module instead of a word lattice, we will lower the False Acceptance rate but increase the False Rejection one. By adding the *M\_freq* models we will better recognize the most

current requests but with a negative impact on atypical ones. And finally adding local context gives a boost on *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 off-line on large audio databases in order to accurately extract relevant categories of spoken utterances. In the study presented in this paper we have decided to run simultaneously four strategies, according to the ASR output used (word 1-best or word lattice) and the kind of model involved:

1.  $S_1 = 1\text{-best} + M_{\text{init}}$
2.  $S_2 = \text{lattice} + M_{\text{init}}$
3.  $S_3 = \text{lattice} + M_{\text{init}} + M_{\text{freq}}$
4.  $S_4 = \text{lattice} + M_{\text{init}} + M_{\text{freq}} + M_{\text{local}}$

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 and use this information for several purposes. For decision strategy: can agreement situations be used in order to choose a hypothesis  $H_i$  among the four hypothesis  $H_1$ ,  $H_2$ ,  $H_3$  and  $H_4$ ?

## 6 Evaluation

The training corpus used to train the  $M_{\text{freq}}$  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 semantic model is made of 400 different concepts leading to 2030 possible structured interpretations. The average Word Error Rate of the 1-best word string produced by the ASR module is 38%. The test corpus manually annotated within the LUNA project is made of 4.5k utterances containing 7.6 k occurrences of concepts. The four strategies  $S_1 \dots S_4$  have been applied to the ASR output of each message of this test corpus, producing the four interpretation hypotheses  $H_1 \dots H_4$ .

Table 1 presents the Interpretation Error Rates (False Acceptance, False Rejection, Substitution and total IER) according to different decision processes: *baseline* consists in choosing the best interpretation from the ASR 1-best with the  $M_{\text{init}}$  models (namely  $H_1$ ); *agree= $n$*  consists of choosing the interpretation on which at least  $n$  hypotheses (among  $H_1 \dots H_4$ ) agree. If the criterion is not satisfied for a given agreement level (for instance if the 4 hypotheses are different in the *agree=2* situation, the utterance is rejected. An interpretation is considered as correct if the entire frame and all the frame elements (or concepts) are correct.

Selection	<i>baseline</i>	<i>agree=2</i>	<i>agree=3</i>	<i>agree=4</i>
<i>False Accept.</i>	10.1	9.7	8.0	<b>6.9</b>
<i>False Reject.</i>	<b>3.3</b>	3.7	4.6	8.2
<i>Substitution</i>	8.6	8.3	7.5	<b>4.9</b>
<i>Total IER</i>	22.0	21.7	20.1	<b>20.0</b>

TABLE 1: Error rates of the decision strategy according to the different agreement situations

*agree=3* and *agree=4* lead to similar overall IER but to different operating points. The choice between one or the other agreement situation should be guided by the applicative requirements (in favour of less *False Acceptance* or less *False Rejection*).

## 7 Conclusion

This paper presents the Spoken Language Understanding approach explored in the LUNA european project. It addresses two of the main objectives of the project: tight ASR/SLU and SLU/DM coupling and adaptive SLU strategies.

The implemented strategies aim at exploiting the complementarity of several decoding views, each taking into account different levels of contextual information. Beyond the overall improvement that can be obtained in terms of Interpretation Error Rate when exploiting the agreement between several decoding hypothesis, it has been shown that different behaviours of the overall system can be obtained depending on the agreement situations between these decoding views.

## References

- Collin F. BAKER, Charles J. FILLMORE, and John B. LOWE (1998), The Berkeley FrameNet Project, in *Proceedings of the 17th international conference on Computational linguistics*, pp. 86–90, Association for Computational Linguistics, Morristown, NJ, USA, doi:<http://dx.doi.org/10.3115/980845.980860>.
- Geraldine DAMNATI, Frederic BECHET, and Renato DE MORI (2007), Spoken Language Understanding strategies on the France Telecom 3000 Voice Agency corpus, in *IEEE ICASSP*, Honolulu, HI.
- A. L. GORIN, G. RICCARDI, and J.H. WRIGHT (1997), How May I Help You ?, in *Speech Communication*, volume 23, pp. 113–127.
- Mehryar MOHRI, Fernando PEREIRA, and Michael RILEY (1997), AT&T FSM Library - Finite State Machine Library, *AT&T Labs - Research*, URL <http://www.research.att.com/sw/tools/fsm/>.
- Christian RAYMOND, Giuseppe RICCARDI, Kepa Joseba RODRIGUEZ, and Joanna WISNIEWSKA (2007), The LUNA Corpus: an Annotation Scheme for a Multi-domain Multilingual Dialogue Corpus, in *The 11th Workshop on the Semantic and Pragmatics of Dialogue DECALOG'07*.

