Sequential Decision Strategies for Machine Interpretation of Speech

Christian Raymond, *Associate Member, IEEE*, Frédéric Béchet, *Member, IEEE*, Nathalie Camelin, *Student Member, IEEE*, Renato De Mori, *Fellow, IEEE*, and Géraldine Damnati, *Member, IEEE*

***Abstract*—Recognition errors made by automatic speech recognition (ASR) systems may not prevent the development of useful dialogue applications if the interpretation strategy has an introspection capability for evaluating the reliability of the results. This paper proposes an interpretation strategy which is particularly effective when applications are developed with a training corpus of moderate size. From the lattice of word hypotheses generated by an ASR system, a short list of conceptual structures is obtained with a set of finite state machines (FSM). Interpretation or a rejection decision is then performed by a tree-based strategy. The nodes of the tree correspond to elaboration-decision units containing a redundant set of classifiers. A decision tree based and two large margin classifiers are trained with a development set to become interpretation knowledge sources. Discriminative training of the classifiers selects linguistic and confidence-based features for contributing to a cooperative assessment of the reliability of an interpretation. Such an assessment leads to the definition of a limited number of reliability states. The probability that a proposed interpretation is correct is provided by its reliability state and transmitted to the dialogue manager. Experimental results are presented for a telephone service application.**

***Index Terms*—Confidence measures, decision strategy, speech recognition, spoken dialogue systems, spoken language understanding.**

I. Introduction

IN SPOKEN dialogues for telephone applications, interpretation consists in finding instances of conceptual structures representing knowledge in the domain semantics. Spoken language understanding (SLU) is the process of obtaining interpretations which are structures of semantic constituents. Interpretation is concept recognition and can be seen as a problem-solving activity based on features extracted from sentences. This type of problem solving is not necessarily limited to parsing under the control of a context-free grammar. There are many motivations supporting this consideration. Semantic knowledge can be context-sensitive. Given a sentence made of a sequence of words, it is possible that not all the words in the sequence are relevant for the expression of concepts in the domain. The same word can be essential for hypothesizing more than one conceptual constituent. As opposed to parsing, generation of concept hypotheses can be successful even if a sentence is not completely generated by a grammar.

In a problem solving perspective, redundant semantic knowledge sources (KS) can be used for improving interpretation accuracy. Each KS represents a different view point for interpretation, considering different words and contexts as essential or useful for representing a concept. For example, a set of different classifiers, with totally or partially different features, can be used in conjunction with grammars relating words with conceptual constituents and their structures. Furthermore, successive refinements can be performed in a sequential interpretation strategy in order to improve accuracy.

Various formalisms have been proposed for describing semantic structures. Essentially they are all based on entities and relations. Let us call semantic constituent an instance of an entity whose presence can be directly hypothesized by transducers or classifiers which take only words and parts of speech (POS) tags as input. These constituents can be combined for obtaining more complex conceptual structures. A formal theory of composition can be found, for example, in [1]. The theory shows how compositions of objects into structures are performed using context-sensitive rules. Some rules represent actions by functions and arguments. Arguments are semantic objects. Description languages like OWL of w3c (www.w3c.org) have been recently introduced to describe such semantic structures.

The focus of this paper is on the detection of semantic constituents and on the evaluation of their reliability. In the attempt to reduce the effect of recognition errors and knowledge imprecision, a sequential decision strategy is proposed. The strategy starts with the generation of hypotheses about elementary semantic constituents, also called concept tags, from a lattice of word hypotheses produced by an automatic speech recognition (ASR) system. Along the line of solutions proposed in [2], [3] a finite state machine (FSM) transducer is introduced for translating patterns of words and POS into a concept tag. Details of this approach are given in [4] and will be briefly summarized in Section II.

Once an interpretation has been hypothesized, the next step in the strategy consists of estimating its confidence. Two basic approaches have been proposed for estimating the confidence of an interpretation. One computes the posterior probability of an interpretation from the posterior probabilities, obtained with

Manuscript received April 29, 2005; revised January 6, 2006. This work was supported by France Télécom R&D under Contract 021B178. Part of the research described here was performed in the framework of the European Network of Excellence PASCAL. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dilek Hakkani-Tur.

C. Raymond, F. Béchet, and R. De Mori are with the Computer Laboratory (LIA), University of Avignon, 84911 Avignon Cedex 09, France (e-mail: christian.raymond@univ-avignon.fr; frederic.bechet@univ-avignon.fr; renato.demori@univ-avignon.fr).

N. Camelin is with the Computer Laboratory (LIA), University of Avignon, 84911 Avignon Cedex 09, France and also with the Speech Group, France Télécom R&D, 22307 Lannion Cedex 07, France (e-mail: nathalie.camelin@univ-avignon.fr).

G. Damnati is with the Speech and Sound Technologies and Processing Laboratory, France Télécom R&D, 22307 Lannion Cedex 07, France (e-mail: geraldine.damnati@rd.francetelecom.com).

Digital Object Identifier 10.1109/TASL.2006.876862

acoustic and language models, of words supporting the interpretation [5]. The other computes a posterior probability from scores obtained by a set of features related to various levels in the decoding process, including acoustic, linguistic, and semantic information as well as dialogue context [6]–[9].

The confidence estimation process proposed in this paper for the validation of semantic hypotheses is based on a sequential strategy. It is represented by a decision tree whose nodes are decision units. At the root of the tree, decision is made based on a semantic confidence established by the agreement of a redundant set of classifiers trained for confirming the hypotheses generated by FSMs. Classifiers and FSMs evaluate interpretations from different view points. In general, classifiers are trained from labeled examples and make decisions based on features which are automatically selected, while devices using FSMs, essentially evaluate word sequences using prior linguistic knowledge.

In contrast to the use of classifiers for semantic chunking [10], in the approach proposed here, automatically trained classifiers are used to validate concept tag hypotheses generated with conceptual language models (LMs). In contrast with [11], the semantic confidence for a concept tag is not obtained by composing confidences of words, but by a direct, multiple-view, global evaluation of the presence of a concept in a sentence.

For many SLU applications, it is not essential to compute the probability of an interpretation, but to find conditions for which what is found is highly likely to be correct. The dialogue strategy may decide to ask the user for a confirmation or a clarification, depending on whether or not the correctness probability is high. A reliability state depending on conditions on confidence indicators is assigned to each concept hypothesis. The number of reliability states is deliberately small in order to guarantee that accurate correctness probabilities can be estimated for each state.

After presenting some related work on confidence measures in Section III, Section IV describes this general strategy. The classifiers used in this study as well as the confidence features are presented in Section V. The decision strategy is evaluated on a telephone service corpus provided by France Telecom R&D. These experiments are detailed in Section VI. Finally, an error correction and rejection strategy is presented in Section VII.

II. Generation of a Structured n-Best List of Conceptual Interpretations

Interpretation starts with a translation process in which stochastic LMs are implemented by FSMs which output labels for semantic constituents. These semantic constituents belong to a small set of major ontological categories (such as `thing`, `event`, `state`, `action`, `place`, `path`, `property`, `amount`, ...). In this paper, these semantic constituents are called *concept tags* and are noted γ. To each concept tag γ is attached the word string γ^w supporting the concept and from which the concept value (e.g., date or numerical information) can be extracted. For example, to the concept tag $\gamma = LOCATION$ can be associated the word string $\gamma^w =$ "*station de métro de la place de l'Opéra*" (*Opera subway station*) which leads to the concept value `METRO:OPERA`. The interpretation of an utterance containing L concepts is represented by both a *concept tag sequence* (noted $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_L\}$) and the corresponding concept word string sequence supporting each tag (noted $\Gamma^w = \{\gamma_1^w, \gamma_2^w, \ldots, \gamma_L^w\}$). Notice that supports of different tags may share some words.

There is an FSM for each elementary conceptual constituent. Each FSM implements a finite-state approximation of a natural language grammar. These FSMs are transducers that take words at the input and output the concept tag conveyed by the accepted phrase. Their definitions depend on the dialogue strategy. FSMs can be either related to dialogue management (confirmation, contestation, ...) or to the application domain (location, date, ...). They can be manually written for domain-independent conceptual constituents (e.g., dates or amounts), or data-induced when enough training data is available. All these transducers are grouped together into a single transducer, called T_{Concept}, which is the union of all of them.

An interpretation activity leading to the output of an n-best list of concept sequences is presented in detail in [4] and can be summarized as follows.

- A first ASR module generates a word graph (G_W) of word hypotheses by means of a generalist LM.
- G_W is composed with the transducer T_{Concept} and the result of this composition is the transducer T_{WC} (a path in T_{WC} is either a word string if one keeps only the input symbols or a concept tag string if one considers the output symbols of the transducer).
- The n-best list of concept tag sequences $\Gamma_1, \Gamma_2, \ldots, \Gamma_n$ is obtained by enumerating the n-best paths on the output symbols of T_{WC}.
- To each sequence Γ_i is attached a word graph G_{Γ_i} which is the set of paths in T_{WC} that output Γ_i.
- By enumerating the m-best paths in G_{Γ_i}, one obtains the best word strings $W_{i,j}$ (with $1 \leq j \leq m$) supporting Γ_i.
- All the filler words (words that do not belong to the support of any concept tag) are removed from the m word strings $W_{i,j}$ in order to produce the m concept word string sequences supporting $\Gamma_i : \Gamma_{i,1}^w, \Gamma_{i,2}^w, \ldots, \Gamma_{i,m}^w$.
- Finally, only the m' (with $m' \leq m$) unique concept word string sequences are kept (the word strings differing only because of filler words are grouped together).

All the operations presented on the FSMs are made with the AT&T FSM toolkit [12].

The result of the translation process is a *Structured N-Best* list of interpretations called $S_{n\text{best}}$. The information is structured according to three levels:

1) first level: the n-best list of concept tag sequences $\Gamma_1, \Gamma_2, \ldots, \Gamma_n$;
2) second level: the m-best list of concept word string sequences $\Gamma_{i,1}^w, \Gamma_{i,2}^w, \ldots, \Gamma_{i,m}^w$ for each interpretation Γ_i;
3) third level: the best word string $W_{i,j}$ found in G_W for each concept word string sequence $\Gamma_{i,j}^w$.

Each element of $S_{n\text{best}}$ is scored with the posterior probability of the conceptual hypothesis given the acoustic observations.

$S_{n\text{best}}$ can be seen as an abstraction of all the possible interpretations of an utterance, a simple example is given in Table I

TABLE I
EXAMPLE OF STRUCTURED n-BEST LIST $S_{n\text{best}}$ OBTAINED ON A WORD GRAPH CORRESPONDING TO THE UTTERANCE: *pas loin du métro Opéra* (*not far from the Opera subway station*)

rank	interpretation	score
Γ_1	• γ_1 =NEAR γ_2 = LOCATION	**0.90**
$\Gamma^w_{1,1}$	→ γ_1^w =[pas loin] γ_2^w =[metro]	*0.57*
$W_{1,1}$	pas loin du metro euh a (*not far from metro euh a*)	
$\Gamma^w_{1,2}$	→ γ_1^w =[pas loin] γ_2^w =[metro opera]	*0.21*
$W_{1,2}$	pas loin du metro opera (*no far from metro opera*)	
$\Gamma^w_{1,3}$	→ γ_1^w =[pas loin] γ_2^w =[trocadero]	*0.12*
$W_{1,3}$	pas loin du trocadero a (*not far from trocadero a*)	
Γ_2	• γ_1 =NEAR γ_2 = MONEY	**0.10**
$\Gamma^w_{2,1}$	→ γ_1^w =[pas loin] γ_2^w =[vingt euros]	*0.06*
$W_{2,1}$	pas loin de vingt euros a (*not far from twenty euros a*)	
$\Gamma^w_{2,2}$	→ γ_1^w =[pas loin] γ_2^w =[trente euros]	*0.04*
$W_{2,2}$	pas loin de trente euros a (not far from thirty euros a)	

for the utterance *pas loin du métro Opéra* (*not far from the Opera subway station*) with the following concept tags: *NEAR*, *LOCATION*, and *MONEY*.

III. RELATED WORK ON CONFIDENCE MEASURES AND DECISION STRATEGY

Estimating the confidence of an interpretation raises several issues: choosing the span of the confidence measures (word, conceptual constituent or utterance), defining the set of features involved in the confidence estimation, combining efficiently the different features, and choosing a decision strategy that takes into account all the features obtained.

For example, the problem of using recognition confidence scoring for speech understanding has been investigated and a discussion on sentence and word level features can be found in [13]. In [14], two confidence measures are introduced for ASR output. A measure for content words is defined as the sum of posterior probability of sentences in an n-best list containing the word. A measure for a concept category is also introduced as the sum of the products of sentence posterior probabilities times a normalized inverse document frequency on the content words in the sentences in the n-best list containing the concept.

In spoken dialogue systems, it is important to use confidence measures that integrate information related to the whole dialogue context rather than just having features based only on acoustic and language model cues. The dialogue context can be taken into account by defining semantic confidence measures, related to the understanding of an utterance, and by integrating dialog expectations. The integration of dialogue manager expectations is proposed in [15]: a semantic parser processes sentences provided by an ASR decoder and the semantic content of each hypothesis is matched with dialogue predictions and utterance type classification based on prosodic cues. Similarly, in [6], the dialogue expectations are represented by clusters of dialogue prompts that are used as features, in conjunction with acoustic and linguistic features, in a decision tree trained to assign confidence to a semantic interpretation.

For the issue of combining multiple knowledge sources at different levels, in [11], previous approaches to integrate semantic and other ASR features are reviewed and it is noticed that in most of the cases their integration into the decision process is rather ad hoc. In the same paper, both word and concept level confidence measures are considered. Concept hypotheses are associated with nonoverlapping word sequences and the concept confidence is a function of word semantic confidence which in turn may depend on features which can be extracted from the entire sentence.

Most of the previous studies use concept confidence measures for validation or rejection of a concept hypothesis according to a threshold on the confidence value. However, utterance-level confidence measures have also been explored, for example in [15], [16], a *global reliability* is assigned to an interpretation. Such a value can be used in a rejection strategy or for rescoring a set of alternative interpretation hypotheses.

Compared to these previous studies, the strategy proposed in this paper highlights the following key points.

- The confidence in a concept tag hypothesis is given according to two dimensions: confidence in the concept tag γ and confidence in the words γ^w supporting the concept. In addition to confidence measures given at the concept level, a limited set of *reliability states* is defined for characterizing a whole utterance interpretation Γ^w.
- A different set of features is used for each dimension, involving semantic information through semantic classifiers, linguistic and acoustic information and dialogue expectation.
- The various features are combined through classifiers but multiple view points are considered: the consensus among different decision processes defines a level of reliability for a concept or an utterance interpretation.
- The reliability state of a hypothesis corresponds to a global confidence measure. The kind of interpretation errors that are expected in each state can also be predicted. The joint utilization of the states with the alternative hypotheses of $S_{n\text{best}}$ leads to an error correction strategy that can reject or add conceptual constituents to the best interpretation obtained in the first stage of the SLU process.

IV. DECISION PROCESS FOR SEMANTIC INTERPRETATION

In order to assign a reliability state to an interpretation Γ, some redundancies are introduced in the generation of concept tag hypotheses. Redundancy models the fact that an interpretation of a sentence or a discourse is more reliable if different experts using different knowledge and view points arrive at the same conclusion. For this purpose, different types of classifiers have been considered.

Combination of scores at different levels has been proposed, for example, in [9] where different merging methods of several

classification processes are compared. The approach proposed here consists in finding an interpretation strategy that uses different groups of classifiers for performing sequences of decisions. The consensus among different decision processes is used to define a reliability state in which a correctness probability is evaluated.

Interpretation is performed by a decision process based on a diagnostic tree. At a node j of the tree, a decision unit DU_j is applied. The unit performs some computation, evaluates confidence measures or other types of features about the content of $S_{n\mathrm{best}}$, and outputs a decision represented by the truth of a predicate or its negation. This decision is taken according to the level of agreement reached by different classification methods that are run over the same set of features within the unit. A unit makes also available to other units the relevant results of its computation. The diagnostic tree can be automatically trained or manually built (as for the experiments described in the following). Each leaf of the tree corresponds to a reliability state. Decision units are designed to maximize the separation between correct and incorrect samples presented at their input and to provide a good growth in coverage.

Two decision units are introduced in this section: DU_1 which takes as input a word string W and a concept tag sequence Γ and checks the truth of a predicate called $I_c(\Gamma)$; DU_2 which takes a concept word string sequence Γ^w and a set of ASR confidence measures as input and checks the truth of a predicate called $S_c(\Gamma^w)$. These predicates are defined as follows.

- Predicate $I_c(\Gamma)$ is true when all the tags detected in Γ are equally predicted by different semantic classifiers, with the underlying assumption that the consensus between classifiers is correlated with the expectation for a tag of being correct.
- Predicate $S_c(\Gamma^w)$ is true when the supports of all the tags in Γ^w are expected to be correct according to several classification processes.

A development set is used in order to verify that these predicates are good indicators that the correctness probability of a sequence of concept tags is high even if the evaluation of the truth does not require an explicit computation of the correctness probability.

A. *Decision Unit* DU_1

Several studies [17], [18] have shown that classification methods, like support vector machines (SVMs) or boosting algorithms (BOOST), can be an efficient way for hypothesizing semantic entities from transcriptions. This approach has two main advantages. First, the amount of human supervision is limited as no keywords or grammars have to be defined in order to characterize a concept. Second, classifiers are more robust to the noise generated by ASR errors and spontaneous speech effects because they rely on sufficient conditions. They may be trained directly from ASR output and therefore model this noise. In addition to two large-margin classifiers respectively based on *SVM*s and boosting algorithm (*BOOST*), decision-tree based classifiers (semantic classification trees *SCT* [3]) have been found to be useful for concept tag hypothesis validation. In this study, three text classification tools (*SVM*, *BOOST*, and *SCT*), presented in detail in Section V-A, are used in order to give a binary decision on the occurrence of a concept tag ct in a word string W.

The three classifiers receive as their input the words of W together with their POS. Because these tools are based on different classification algorithms with different input formats (bag-of-words or word-strings, for example), they do not always use the same information in order to characterize a concept tag.

Let $V_{ct} = \{ct_1, ct_2, \ldots, ct_k\}$ be the vocabulary of concept tags. Γ is a sequence of symbols in V_{ct} obtained by means of FSMs as presented in Section II. Let V_Γ be a vector with k components in $\{0, 1\}$ where $V_\Gamma[i] = 1$ iff $ct_i \in \Gamma$ and 0 otherwise. The order in which concept tags occur is not taken into account here but only the occurrence or not of each ct_i in W.

For a given word string W, each classification tool T makes k binary decisions on the k concept tags of V_{ct} and produces a vector V_T, similar to V_Γ, where $V_T[i] = 1$ iff the classifier T makes a positive decision for the occurrence of ct_i in W and 0 otherwise. Three vectors are computed, one for each classifier: V_{SVM}, V_{BOOST}, and V_{SCT}.

The decision unit DU_1 evaluates the truth of predicate I_c in the following way:

$$\begin{aligned}\mathrm{DU}_1 : I_c(\Gamma) = \text{True iff} \quad & \forall i \in [1, k] \\ & V_\Gamma[i] = V_{\mathrm{SVM}}[i] = V_{\mathrm{BOOST}}[i] = V_{\mathrm{SCT}}[i]. \end{aligned} \tag{1}$$

Notice that classifiers look at different features in W. In general, they consider a broader context than the one used by the FSMs for producing Γ. Such a broader context may invalidate a hypothesis generated by an FSM. Classifiers may also hypothesize correct concept tags which are not hypothesized by any FSM because the way of generalizing observed examples is different.

B. *Decision Unit* DU_2

While DU_1 evaluates the reliability of a concept tag sequence Γ in a word string W, DU_2 validates the support Γ^w of Γ. Each concept tag support γ^w is represented by a set of ASR confidence features presented in Section V-B.

Following the notion of consensus between classifiers proposed in DU_1, the same set of classifiers is used now in order to classify each γ^w into *correct* or *incorrect* classes according to the reference labels. The training process of the three classifiers on the development corpus is presented in Section V-B.

Each classifier T provides a truth value for a correctness predicate $C_T(\gamma^w)$. Symbol $C_T(\gamma^w)$ indicates the fact that classifier T has labeled the support γ^w as correct based on the confidence features attached to γ^w.

The predicate $S_c(\gamma^w)$ is defined as follows:

$$S_c(\gamma^w) = C_{\mathrm{SVM}}(\gamma^w) \wedge C_{\mathrm{BOOST}}(\gamma^w) \wedge C_{\mathrm{SCT}}(\gamma^w). \tag{2}$$

For a concept tag sequence $\Gamma = \{\gamma_1, \ldots, \gamma_L\}$ with support $\Gamma^w = \{\gamma_1^w, \ldots, \gamma_L^w\}$, the decision unit DU_2, evaluates the truth of the predicate $S_c(\Gamma^w)$ as follows:

$$\mathrm{DU}_2 : S_c(\Gamma^w) = \wedge_{j=1}^{L} S_c\left(\gamma_j^w\right). \tag{3}$$

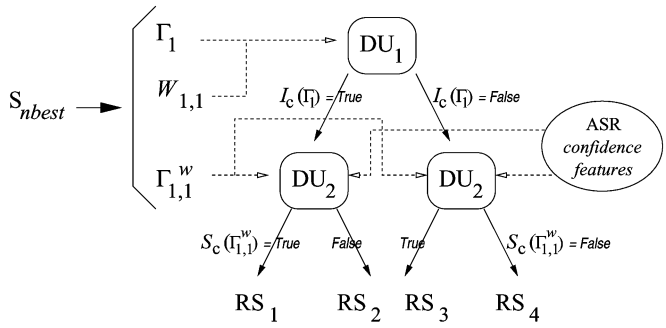

Fig. 1. Interpretation strategy with decision units DU_1 and DU_2.

C. Reliability States

The decision process takes as input a structured n-best list $S_{n\mathrm{best}}$ and starts by processing the top hypothesis Γ_1 with DU_1.

The diagnostic tree implementing the decision process has DU_1 and DU_2 as nodes. The dataflow corresponding to this process is shown in Fig. 1. This first implementation is quite simple; more sophisticated strategies can be defined by considering not only full agreement among the decision processes as in DU_1 and DU_2 but also partial agreements.

Four reliability states are defined, corresponding to the following expressions:

$$\mathrm{RS}_1 : I_c(\Gamma_1) \wedge S_c\left(\Gamma_{1,1}^w\right) \quad \mathrm{RS}_2 : I_c(\Gamma_1) \wedge \overline{S_c\left(\Gamma_{1,1}^w\right)}$$
$$\mathrm{RS}_3 : \overline{I_c(\Gamma_1)} \wedge S_c\left(\Gamma_{1,1}^w\right) \quad \mathrm{RS}_4 : \overline{I_c(\Gamma_1)} \wedge \overline{S_c\left(\Gamma_{1,1}^w\right)}.$$

The probability that an interpretation is correct in a reliability state RS_j is evaluated with a development set together with the percentage of examples (coverage) observed in this state. Each reliability state RS_j has a coverage c_j. To each state is associated an interpretation error probability $P_e(\mathrm{RS}_j)$ estimated with a development set. Let us assume, for example, that an instance $obj(Place_Z)$ of an object *Place* is in reliability state RS_j, then its correctness probability is: $P_c\{obj(Place_Z)\} = 1 - P_e(\mathrm{RS}_j)$.

In this way, complex and, perhaps, rare instances of semantic structures have a correctness probability which has been estimated on a limited number of reliability states.

The normalized cross entropy (NCE) measure as recommended by [9] is used for evaluating the performance of a chain of *DUs*. This measure is based on the mutual information (cross entropy) between the correctness of the concepts output by the system and the confidence scores attached to them, normalized by the maximum cross entropy. NCE is defined as follows:

$$\mathrm{NCE} = \frac{H_{\max} + \sum_{cor.\gamma^w} \log_2(P_c(\gamma^w)) + \sum_{incor.\gamma^w} \log_2(P_e(\gamma^w))}{H_{\max}} \tag{4}$$

with $H_{\max} = -n\log_2(n/N) - (N-n)\log_2(1-(n/N))$, n is the number of correct hypotheses γ^w, n is the total number of hypotheses γ^w, and P_c is the confidence score of the reliability state attached to γ^w.

The diagnostic tree is built by sequentially introducing *DUs* that maximize NCE for the applicable data in the development set.

V. Confidence Features

A. Semantic Classifiers for DU_1

Text classification tools may differ by the classification method used and by the features chosen for representing textual information (word, Part-Of-Speech tag, lemma, stemma, bag of tokens, bag of n-grams, utterance length, etc.). Because there is no generic method that is proven to be the best in all the cases, various classifiers and textual features are combined in this study in order to assign a confidence measure to a sentence interpretation.

The classifiers used in this study are now briefly described.

1) *LIA_SCT* [20] is a classifier based on the SCTs described in [3] for the ATIS task. This classifier takes as input strings of tokens (possibly descriptions at different levels of abstraction, like words and POS tags for example) and dynamically builds regular expressions that span strings of various lengths.
2) *BoosTexter* [18] is a classifier based on a boosting method of *weak* classifiers. The features chosen in our experiments are 1-gram, 2-gram, and 3-gram words and POS tags.
3) *SVM-Torch* [21] is an SVM-based classifier where the input data is a numerical feature vector. In our experiments, all utterances are represented by a *bag of token*, each token being a word or a POS tag.

Because these tools are based on different classification algorithms with different input formats, they do not always use the same information in order to characterize a concept and therefore they react differently to the noise that may affect the ASR results. It is thus expected that agreement or disagreement among classifiers is a useful confidence feature.

The corpus used to train these classifiers is made of user utterances (both manual and ASR transcriptions) where each utterance is labeled with its conceptual interpretation represented by a sequence of concept tags. The classifiers are trained in order to detect the occurrence of each concept tag in an utterance. For the classifiers that accept multilabel samples, only one model is trained with all the concept labels. For those that can handle only one label for each sample, n binary classifiers are trained for the n concept labels of the application targeted.

B. Confidence Classifiers for DU_2

The decision unit DU_2 evaluates the relevance of the word strings γ^w supporting the concept tags with a set of confidence measures given at the concept or the utterance level. These measures are now introduced.

- AC, a descriptor of acoustic confidence [22], is attached to each word string γ^w.
- LC, a descriptor of linguistic confidence inspired by measures proposed in [23], is the ratio, for a given word string candidate, between the number of trigrams observed in the training corpus of the LM versus the total number of trigrams in the same word string. This measure is evaluated

on the best word string W containing γ^w and covering the whole utterance.
- SC, a descriptor of semantic confidence, is the classification scores given to γ by the three classifiers in DU_1.
- R, the rank of the best hypothesis in the n-best list $S_{n\mathrm{best}}$ containing γ^w, based on the likelihood computed by the speech recognition module.
- DC, a descriptor of dialogue context confidence. The dialogue context is represented by the system prompt played before each user's turn. Each prompt is labeled with a tag corresponding to the kind of message given to the user (open prompt, confirmation, specific request, ...). An *a priori* distribution of all the concept tags γ for each prompt label is obtained on the training corpus.

The three classifiers SVM, BOOST, and SCT are trained to classify examples into *correct* or *incorrect* classes thanks to these confidence features. A corpus of examples is built from a set of $S_{n\mathrm{best}}$ n-best lists obtained on a development corpus of utterances. Each concept hypothesis γ^w found in these lists is described by the confidence features introduced above and constitutes a training example for the classifiers. The tag *correct* or *incorrect* is given according to the reference version of the development corpus.

During the decision process of an utterance, the support $\Gamma^w_{1,1} = \{\gamma^w_1, \gamma^w_1, \ldots, \gamma^w_L\}$ of the concept tag sequence Γ_1 is evaluated by DU_2 as follows.
- To each tag support γ^w_i with $i \in [1, L]$ is attached the set of previously described features (AC, LC, R, DC).
- Three classification models (SVM, BOOST, and SCT) are applied to each γ^w_i and output three classification scores for the class *correct*.
- Each score can be seen as a distance between the *correct* and *incorrect* classes and a classification confidence score is then calculated by applying a sigmoid function to this distance.
- If this score is above a given threshold, the predicate $C_T(\gamma^w_i)$ with $T \in \{SVM, BOOST, SCT\}$ is asserted to be true.
- Finally, DU_2 generates a result as presented in Section IV-B.

The classification confidence scores given by the three classification models are also combined and used for a rejection strategy. It will be shown in Section VI that this confidence score is a very powerful feature for rejecting a concept tag hypothesized by the understanding module, if the confidence value obtained is below a given threshold. By tuning this threshold, one can efficiently adjust the recall/precision performance in the concept tag detection of the decision process.

VI. Experiments

A. Experimental Setup

Experiments were carried out on a dialogue corpus provided by France Telecom R&D and collected for a tourism telephone service. The task has a vocabulary of 2200 words and a vocabulary (V_{ct}) of 59 concept tags.

This corpus is divided into three sets: a training corpus (*TRAIN*) containing 13 K utterances manually transcribed and labeled, a development corpus (*DEV*) made of 4 K utterances, and a test corpus (*TEST*) containing 1.5 K utterances both with the manual transcriptions and labeling as well as the corresponding structured n-best lists $S_{n\mathrm{best}}$ output by the SLU module. The amount of concept tags contained in the *TRAIN*, *DEV*, and *TEST* corpora are, respectively, 27 K, 8 K, and 3 K concept tags. The word error rates (WERs) on the development and test corpora, considering the word sequence $W_{1,1}$ in the lists $S_{n\mathrm{best}}$, are 25.8% and 27.0%, respectively.

The FSMs representing the concept tags, as presented in Section II, are trained with the *TRAIN* corpus. The classifiers for DU_1 are trained on the manual transcription of the *TRAIN* corpus as well as the automatic transcription (best word string $W_{1,1}$) of the *DEV* corpus. The automatic transcription is used in order to introduce noise into the training process of the classification models. Indeed, the first decision unit depends on the agreement between the decision processes, not on the individual performance of each classifier. By introducing ASR errors in the training process, the classifiers have to use more contextual information in order to model the concept tags as the word supports of the tags can be incorrect. Therefore, the classifiers implement different viewpoints compared to the FSMs that focus only on the word support of the concept tags. This assessment is evaluated in the following subsection. The classifiers for DU_2 are all trained on the *DEV* corpus.

The results are given according to two measures.
- The *Concept Error Rate* (CER), which is similar to the WER but at the concept level. A conceptual constituent is considered correct only if both its concept tag and its concept value are correct.
- The *coverage* (Cover), which indicates the percentage of utterances accepted by a given decision unit DU_i with respect to the whole number of utterances in the corpus.

B. Evaluation of the Decision Units DU_1 and DU_2

The two decision units DU_1 and DU_2, presented in Sections IV-A and IV-B are considered in this experiment. Table II reports results, in terms of generation of concept tag hypotheses in $\Gamma^w_{1,1}$, before and after each decision unit. These results indicate that the agreement rule among the classifiers is a powerful semantic confidence measure as the CER is clearly linked to the consensus situation reached. The diagnostic tree shown in Fig. 1 is used to define four *reliability states* attached to each leaf node of this tree. The reliability of each state $\mathrm{RS}_{1,2,3,4}$ is estimated on the *DEV* corpus and they can be sorted according to their correctness probability $P_c(RS_i) : P_c(RS_1) > P_c(RS_3) > P_c(RS_2) > P_c(RS_4)$.

These correctness probabilities can be used as confidence scores in order to estimate the NCE value of a given strategy (or a given diagnostic tree). Indeed, the NCE values displayed in Table II validate the diagnostic tree proposed: the highest value (0.27) is obtained for RS_1 by using the four states $\mathrm{RS}_{1,2,3,4}$. In a two-states strategy, the decision unit DU_2 should be preferred to DU_1. As already mentioned, more elaborate strategies with partial agreement among the classifiers can be considered, using the NCE measure as an optimization criterion.

TABLE II
INTERPRETATION ERROR RATES ON THE *TEST* CORPUS FOR $\Gamma^w_{1,1}$ WITH DECISION UNITS DU_1 AND DU2. THE RELIABILITY STATES RS_i ARE INDICATED AS WELL AS THE NORMALIZED CROSS ENTROPY (NCE) OBTAINED BY USING THE CONFIDENCE MEASURE ATTACHED TO EACH RS_i AT THE CONCEPT LEVEL

Condition	*All*	DU_1	$\overline{DU_1}$	DU_2	$\overline{DU_2}$	$DU_1 \wedge DU_2$	$DU_1 \wedge \overline{DU_2}$	$\overline{DU_1} \wedge DU_2$	$\overline{DU_1} \wedge \overline{DU_2}$
Cover	*100%*	74.6%	25.4%	67.1%	32.9	58.4%	16.2%	7.6%	17.7%
CER	17.0	11.4	27.4	7.9	30.2	5.9	28.6	18.0	30.8
RS	-	-	-	-	-	RS_1	RS_2	RS_3	RS_4
NCE	-	0.03	0.04	0.17	0.08	**0.27**	0.09	0.00	0.07

TABLE III
PERFORMANCE OF THE DECISION UNITS DU_1 AND DU_2 ACCORDING TO THE NUMBER OF CLASSIFIER INVOLVED

DU	**class**	**coverage**	**CER**
none	*none*	100%	17.0
DU_1	**+SCT**	86.5%	14.6
	+BOOST	76.1%	11.5
	+SVM	74.6%	11.4
$+DU_2$	**+SCT**	65.6%	9.1
	+BOOST	61.5%	6.9
	+SVM	58.4%	5.9

The text classifiers used in DU_1 are trained on the ASR output of the development corpus leading to a CER of 11.4 for a coverage of 74.6%. If all the decision processes (FSMs and classifiers) were learned only on the reference transcriptions, a CER value of 13.8 would be achieved with a coverage of 83%. As expected, when they are trained on clean data, the classifiers are likely to focus more on the supports of the concept tags rather than on their contexts of occurrence. Therefore, the agreement among the decision processes is 10% higher with the training on manual transcriptions but the CER is significantly worse (from 11.4 to 13.8). Since it is the identification of the most reliable hypotheses that is the goal of each decision unit, it is the training with ASR output that is chosen in the strategy presented here.

The gain obtained by using three different classification methods in DU_1 and DU_2 is illustrated in Table III. The order in which the classifiers are used is not relevant. As it is shown, each classifier added leads to a decrease in the error rates. However, the gain obtained by adding the third classifier is rather small, justifying the use of no more than three classification methods.

VII. ERROR CORRECTION AND REJECTION STRATEGY

The four reliability states can be interpreted as follows.

- RS_1 contains reliable interpretations, with a low CER, the remaining errors being concepts not detected by either the FSMs or the text classifiers, and therefore very likely to be deletion errors.
- RS_2 contains interpretations validated by DU_1 but not DU_2; therefore, if the concept tags are likely to be correct, some of their values might be erroneous, leading to substitution errors.
- In RS_3, there is no total agreement on the concept tag sequence obtained with the FSMs as one or more classifiers have hypothesized different concept tags from those occurring in Γ_1; however, because all the supports of the concepts of $\Gamma^w_{1,1}$ are validated by DU_2, it is very likely that deletion errors are predominant.

TABLE IV
ERROR DISTRIBUTION ACCORDING TO THE RELIABILITY STATE *RS* AND THE CORPUS (*DEV* OR *TEST*)

Errors	**deletion**		**insertion**		**substitution**	
corpus	*dev*	*test*	*dev*	*test*	*dev*	*test*
ALL	34.2%	34.4%	22.4%	21.6%	**43.5%**	**44.0%**
RS_1	**84.0%**	**54.8%**	6.7%	10.8%	9.2%	34.4%
RS_2	10.3%	31.9%	21.7%	18.0%	**68.0%**	**50.0%**
RS_3	**99.5%**	**63.5%**	0.0%	5.8%	0.5%	30.8%
RS_4	13.4%	22.2%	31.9%	31.0%	**54.8%**	**46.8%**

TABLE V
AVERAGE *Oracle CER* IN THE LISTS OF HYPOTHESES ATTACHED TO EACH UTTERANCE WITH THE AVERAGE MINIMUM NUMBER OF HYPOTHESIS (n) THAT HAS TO BE KEPT TO REACH THIS CER IN BOTH THE STANDARD AND THE STRUCTURED n-BEST LISTS OF HYPOTHESES

n-best	RS_1	RS_2	RS_3	RS_4
Oracle CER	3.0	16.0	9.2	17.8
n in standard n-best	15	14	23	44
n in S_{nbest}	4	5	5	5

- In RS_4, there is no agreement found in DU_1 or DU_2, this state has the highest CER with all types of errors (deletion, insertion, substitution).

Table IV summarizes the distribution of the understanding errors according to the reliability states on the *DEV* and *TEST* corpora. As one can see, the hypotheses made about the main source of errors of each state *RS* are very well validated on the *DEV* corpus (on which the classifiers have been trained), and in a smaller proportion on the *TEST* corpus. The goal of the error correction and rejection strategy is to take advantage of this knowledge in order to either look for a possible correction of $\Gamma^w_{1,1}$ in the structured n-best lists $S_{n\mathrm{best}}$ or reject some concepts $\gamma^w \in \Gamma^w_{1,1}$ or even reject the interpretation of the whole utterance. Furthermore, the way a possible correction is looked for in $S_{n\mathrm{best}}$ can be made dependent on the *RS* state associated to $\Gamma^w_{1,1}$.

For evaluating the potential of the error correction process, it is interesting to estimate the lower bound of CER that can be found in the n-best lists of hypotheses. This lower bound is called the *Oracle CER*. It can be obtained from a list of hypotheses by selecting the one with the lowest CER according to the reference. The Oracle CER measures are given in Table V according to each reliability state $\mathrm{RS}_{1,2,3,4}$, for two kinds of list of hypotheses: the standard n-best lists output by the ASR module and the $S_{n\mathrm{best}}$ lists. As we can see, the $S_{n\mathrm{best}}$ lists outperform significantly the standard n-best lists: by keeping the top five hypotheses, the Oracle CER is reached for the four states. It is interesting to point out that the Oracle CER is well correlated with the reliability states: from about 3% in the high reliability state RS_1 up to 18% for the low reliability state RS_4.

A. Error Correction Strategy

The error correction strategy is dependent on the reliability state, and is made of the following steps.

1) *Reliability state* RS_1: Because interpretation hypotheses are very reliable in this state, no correction is applied to $\Gamma^w_{1,1}$. The additional concepts that can be found in the other hypotheses of $S_{n\mathrm{best}}$ are also sent to the dialogue manager, for possible insertion in Γ_1, and on which confirmation by the user is needed. By adding these concepts the recall measure on the concept extraction of this state increases from 94.7% to 96.5%.
2) *Reliability state* RS_2: The major source of errors of this state being substitution errors, the alternative values to the concepts of $\Gamma^w_{1,1}$ found in $S_{n\mathrm{best}}$ are also sent to the dialogue manager. The dialogue context can then be used to filter these lists of values. The recall measure increases from 76.7% to 82.7% by adding these concept values.
3) *Reliability state* RS_3: This state is very likely to contain a lot of deletion errors. In order to correct some of them, linguistic inconsistencies have been found using an explanation-based learning approach. Examples of inconsistencies were searched for in the development set. Each example was manually generalized to derive a pattern for detecting inconsistency and a corresponding pattern to represent the correction. If an inconsistency pattern is found in $\Gamma^w_{1,1}$ and the corresponding correction pattern is found in $S_{n\mathrm{best}}$, then correction is applied. For example, the following patterns have been considered for correcting deletions in RS_3 due to an ASR problem on begin and end point detection: *missing verb at the beginning of an utterance* and *addition of a concept with value at the end of an utterance*. These types of corrections were applied to the test samples in this state and on the 45 utterances with deletion errors found, 12 of them have been corrected by these rules.
4) *Reliability state* RS_4: For the hypotheses falling in this low reliability state, the decision-tree based error correction strategy presented in [24] is applied. This strategy trains a decision tree on the development corpus to accept or reject a correction of $\Gamma^w_{1,1}$ found in $S_{n\mathrm{best}}$, thanks to the set of confidence features used in the decision unit DU_2. This strategy applied to RS_4 decreases the CER from 30.8 to 28.9 on the *TEST* corpus. Because of the low reliability of the hypotheses attached to this state, a rejection rule can also be applied, leading the dialogue manager to ask for a repetition. By applying a threshold on the confidence score produced by merging the three classification scores in the decision unit DU_2, some utterances can be discarded. As an example, by discarding 40% of the utterances of RS_4, the CER drops from 30.8 to 24.4 and by keeping only 30% of them, the CER reaches 20.5%.

B. Concept Rejection Strategy

By using a threshold on the confidence score estimated in the decision unit DU_2, as presented in Section V-B, one can reject the concepts having a score below this threshold. This rejection strategy is compared to a baseline strategy based on the acoustic confidence scores only. The results are presented in Fig. 2 with the *False Acceptance* $(\mathbf{FA})$ versus *Correct Acceptance* $(\mathbf{CA})$ curves for the four reliability states on the *TEST* corpus. These measures are defined as follows:

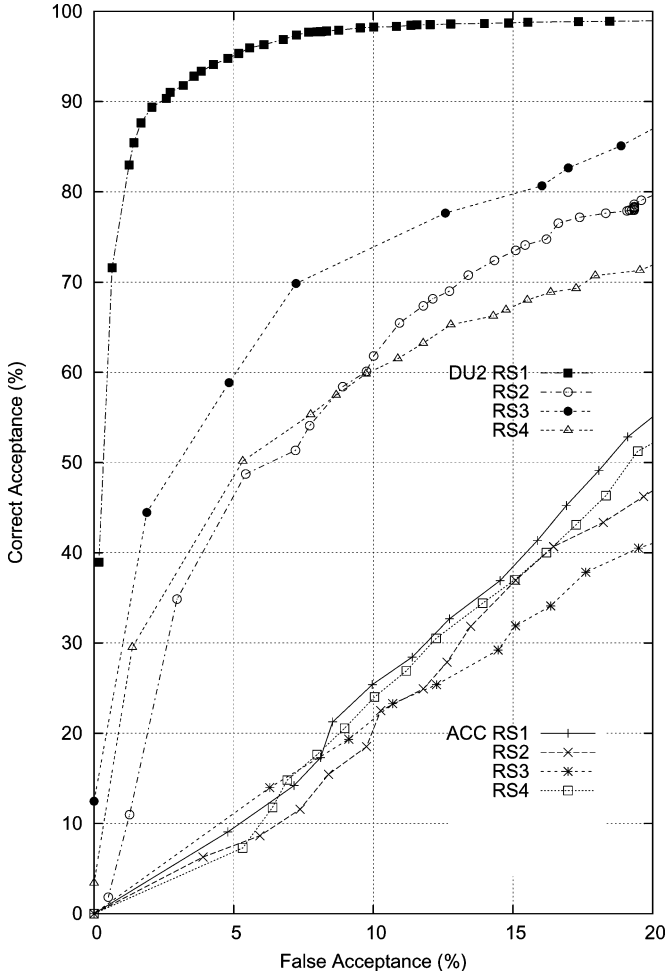

Fig. 2. *False Acceptance* $(\mathbf{FA})$ versus *Correct Acceptance* $(\mathbf{CA})$ curves for two confidence measures, one based on the acoustic confidence (ACC) and one based on the classifier scores obtained in the decision unit DU_2, for the four reliability states $\mathrm{RS}_{1,2,3,4}$.

$$\mathbf{FA} = \frac{\#\text{ of falsely accepted concepts}}{\text{Total \# of negative examples}} \times 100$$

$$\mathbf{CA} = \frac{\#\text{ of correctly accepted concepts}}{\text{Total \# of positive examples}} \times 100.$$

The strategy using the DU_2 score significantly outperforms the baseline strategy. At a 5% FA operating point, 95% of the correct concepts of RS_1 are kept (and this state contains nearly 60% of the utterances), 60% for RS_2, and about 50% for RS_3 and RS_4. This measure can be used by the dialogue manager for identifying, in a given interpretation, the concepts that are almost certain from those that should be confirmed by the user in the next dialogue turn.

VIII. Conclusion

A sequential interpretation strategy has been proposed based on a tree of decision units. Strategy design follows the conjec-

ture that the word hypotheses about a spoken sentence have to be interpreted with different types of semantic knowledge. Consensus about an interpretation formulated from the different points of view of the different semantic knowledge sources results in a high probability that the interpretation is correct. Different knowledge sources are obtained with different automatically trained classifiers and FSMs. Classifiers use features which are words, POS tags, degrees of confidence, and agreement. There may be a classifier for each conceptual constituent which is trained to discriminate between a given concept and all the others, including the empty one. Some classifiers are also used in conjunction with inference procedures for performing error correction of the first candidate in the structured n-best list.

With the proposed strategy, the probability that an interpretation is correct does not necessarily rely on the frequency of observation of words expressing a concept, but on the discriminative power of features which are selected during classifier training with a rather small development set.

The sequential interpretation strategy identifies reliability states with associated probabilities of correct interpretation. Experimental evidence has shown that the correctness probability is very high for certain states, suggesting situations in which the dialogue manager may not need to ask for a confirmation. In other states, the interpretation hypotheses are unreliable and improvements can be obtained by error correction. A method for validating each concept hypothesis has been presented. It is shown that a strategy using a score derived from the combination of three classifiers significantly outperforms a strategy merely based on acoustic confidence. The proposed approach is particularly interesting when applications have to be developed with a training corpus of moderate size.

REFERENCES

[1] R. Jackendoff, *Semantic Structures*. Cambridge, MA: MIT Press, 1990.

[2] E. Levin and R. Pieraccini, "Concept-based spontaneous speech understanding system," in *Proc. Eur. Conf. Speech Commun. Technol.*, Madrid, Spain, 1995, pp. 555–558.

[3] R. Kuhn and R. De Mori, "The application of semantic classification trees to natural language understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 5, pp. 449–460, May 1995.

[4] C. Raymond, F. Béchet, R. De Mori, and G. Damnati, "On the use of confidence for statistical decision in dialogue strategies," in *Proc. 5th SIGdial Workshop on Discourse and Dialogue, Association for Comput. Ling.*, M. Strube and C. Sidner, Eds., Cambridge, MA, April 30–May 1, 2004, pp. 102–107.

[5] K. Hacioglu and W. Ward, "A concept graph based confidence measure," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Orlando, FL, 2002, pp. 225–228.

[6] S. Pradhan and W. Ward, "Estimating semantic confidence for spoken dialogue systems," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Orlando, FL, 2002, pp. 233–236.

[7] C. Raymond, F. Béchet, R. De Mori, G. Damnati, and Y. Estève, "Automatic learning of interpretation strategies for spoken dialogue systems," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Montréal, QC, Canada, 2004, vol. 1, pp. 425–128.

[8] R. Sarikaya, Y. Gao, and M. Picheny, "A comparison of rule-based and statistical methods for semantic language modeling and confidence measurement," in *Proc. HLT-NAACL Conf., Short Papers*, Boston, MA, May 2004, pp. 65–68.

[9] D. Hakkani-Tür, G. Tur, G. Riccardi, and H. K. Kim, "Error prediction in spoken dialog: From signal-to-noise ratio to semantic confidence scores," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Philadelphia, PA, 2005, pp. 1041–1044.

[10] K. Hacioglu, "A lightweight semantic chunker based on tagging," in *Proc. HLT-NAACL Conf.*, Boston, MA, May 2004, pp. 145–148.

[11] R. Sarikaya, Y. Gao, M. Picheny, and H. Erdogan, "Semantic confidence measurement for spoken dialog systems," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 534–545, Jul. 2005.

[12] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Comput., Speech, Lang.*, vol. 16, no. 1, pp. 69–88, 2002.

[13] T. J. Hazen, S. Seneff, and J. Polifroni, "Recognition confidence scoring and its use in speech understanding systems," *Comput., Speech, Lang.*, vol. 16, no. 1, pp. 49–68, 2002.

[14] K. Komatani and T. Kawahara, "Generating effective confirmation and guidance using two-level confidence measures for dialogue systems," in *Proc. Int. Conf. Spoken Lang. Process.*, Beijing, China, 2000, vol. 2, pp. 648–651.

[15] S. Abdou and M. Scordilis, "Integrating multiple knowledge sources for improved speech understanding," in *Proc. Eur. Conf. Speech Commun. Technol.*, Aalborg, Denmark, 2001, pp. 1783–1786.

[16] P. Carpenter, C. Jin, D. Wilson, R. Zhang, D. Bohus, and A. Rudnicky, "Is this conversation on track?," in *Proc. Eur. Conf. Speech Commun. Technol.*, Aalborg, Danemark, 2001, vol. 1, pp. 455–458.

[17] P. Haffner, G. Tur, and J. Wright, "Optimizing SVMs for complex call classification," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, Hong-Kong, 2003, pp. 632–635.

[18] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Mach. Learning*, vol. 39, pp. 135–168, 2000.

[19] NIST, The 2001 NIST Hub-5 Evaluation Plan [Online]. Available: http://www.nist,gov/speechrtests/ctr/h5_2001/h5-01vl.l.pdf 2001

[20] F. Béchet, A. Nasr, and F. Genet, "Tagging unknown proper names using decision trees," in *Proc. 38th Annu. Meeting Assoc. Comput. Ling.*, Hong-Kong, China, 2000, pp. 77–84.

[21] R. Collobert, S. Bengio, and J. Mariéthoz, Torch: A Modular Machine Teaming Software Library, Institut Dalle Intelligence Artificielle Perceptive, Martigny, Switzerland, Tech. Rep. IDIAP-RRQ2-46, 2002.

[22] C. Raymond, Y. Estève, F. Béchet, R. De Mori, and G. Damnati, "Belief confirmation in spoken dialogue systems using confidence measures," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, St. Thomas, U.S. Virgin Islands, 2003.

[23] Y. Estève, C. Raymond, R. De Mori, and D. Janiszek, "On the use of linguistic consistency in systems for human-computer dialogs," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 746–756, Nov. 2003.

[24] C. Raymond, F. Béchet, N. Camelin, R. De Mori, and G. Damnati, "Semantic interpretation with error correction," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, Philadelphia, PA, 2005, vol. 1, pp. 29–32.

Christian Raymond (A'06) received the M.S. and Ph.D. degrees in computer science from the University of Avignon, Avignon, France, in 2000 and 2005, respectively.

He is a Researcher in the Computer Laboratory (LIA), University of Avignon. His research activities are focused on computer interpretation of spoken sentences and on the use of semantic knowledge in statistical language modeling.

Frédéric Béchet (M'05) received the Ph.D. degree in computer science from the University of Avignon, Avignon, France, in 1994.

Since 1995, he has been a Assistant Professor and a Researcher at the Computer Laboratory (LIA), University of Avignon. He is the author/coauthor of over 40 refereed papers in journals and international conferences. He was an invited professor at the AT&T Research Lab, Florham Park, NJ, from August 2001 until September 2002. His research activities focus mainly on spoken language understanding in human–computer dialogue context and language models for automatic speech recognition.

Dr. Béchet has served on the scientific committees of several international conferences (ICASSP, Eurospeech, Eusipco, ASRU, HLT/EMNLP,) and has been an invited reviewer for several journals including: *Speech Communication*, the IEEE SIGNAL PROCESSING LETTERS, and the IEEE TRANSACTIONS ON SPEECH AUDIO PROCESSING.

Nathalie Camelin (S'06) received the M.S. degree in computer science from the University of Avignon, Avignon, France, in 2004. She is currently pursuing the Ph.D. degree in both the Computer Laboratory (LIA), University of Avignon, and the Speech Group, France Telecom R&D Laboratory, Lannion, France.

Her research activities are focused on spoken language understanding of telephone speech.

Renato De Mori (M'93–SM'89–F'94) received the doctorate degree in electronic engineering from the Politecnico di Torino, Turin, Italy.

He is the author or editor of four books and has published more than 100 scientific papers on many international journals. His major contributions have been in the area of automatic speech recognition, computer arithmetic, software engineering and human–machine interfaces.

Prof. De Mori was Chief Editor of *Speech Communication* (2003–2005) (member of the editorial board since 1982) and Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (1998–1992). He is on the Editorial Board of the *Computer Speech and Language* (since 1988). He was on the Editorial Board of *Computational Intelligence* (1990–2002), *Pattern Recognition Letters* (1980–2004), and *Signal Processing* (1979–1989). He has been a member of the Executive Advisory Board at the IBM Toronto Laboratory, Scientific Advisor at France Télécom R&D, Chairman of the Computer and Information Systems Committee, Natural Sciences, and Engineering Council of Canada, Vice-President R&D, Centre de Recherche en Informatique de Montréal. He has been a member of the IEEE Speech Technical Committee, the Interdisciplinary Board, Canadian Foundation for Innovation, and is currently a member of the Interdisciplinary Committee for Canadian chairs.

Géraldine Damnati (M'06) graduated from École Nationale Supérieure des Télécommunications, Brest, France, in 1996 and received the Ph.D. degree in computer science from the University of Avignon, Avignon, France, in 2000.

Since 1996, she has been involved in speech recognition research at France Télécom R&D, Lannion, France. She has been particularly involved in language modeling studies and in the design of continuous speech recognition models for advanced interactive vocal services. Her current centers of interest include language modeling, speech understanding, and speech-to-speech translation. She is currently member of the speech recognition team in the Speech and Sound Technologies and Processing (SSTP) Laboratory.