SEMANTIC INTERPRETATION WITH ERROR CORRECTION

Christian Raymond[1], *Frédéric Béchet*[1], *Nathalie Camelin*[1], *Renato De Mori*[1], *Géraldine Damnati*[2],

[1] LIA/CNRS - University of Avignon, BP1228 84911 Avignon cedex 09 France
[2] France Télécom R&D - DIH/IPS/RVA 2 av. Pierre Marzin 22307 Lannion Cedex 07, France
{*christian.raymond,frederic.bechet,nathalie.camelin,renato.demori*}*@univ-avignon.fr*
geraldine.damnati@rd.francetelecom.com

ABSTRACT

This paper presents a semantic interpretation strategy, for Spoken Dialogue Systems, including an error correction process. Semantic interpretations output by the Spoken Understanding module may be incorrect, but some semantic components may be correct. A set of situations will be introduced, describing semantic confidence based on the agreement of semantic interpretations proposed by different classification methods. The interpretation strategy considers, with the highest priority, the validation of the interpretation arising from the most likely sequence of words. If the probability, given by our confidence score model, that this interpretation is not correct is high, then possible corrections of it are considered using the other sequences in the N-best lists of possible interpretations. This strategy is evaluated on a dialogue corpus provided by France Telecom R&D and collected for a tourism telephone service. Significant reduction in understanding error rate are obtained as well as powerful new confidence measures.

1. INTRODUCTION

In a previous paper [1], a method has been proposed for obtaining interpretations of spoken sentences using stochastic finite state transducers (SFST). With these transducers an ordered set of N-best lists is obtained with concept-dependent language models (LM). Each list contains sequences of words which generate the same semantic interpretation but may have different acoustic and linguistic confidence. Semantic interpretations may be incorrect, but some semantic components may be correct. It is then useful for the dialogue manager to have the probability that each interpretation component is corrected. This probability should be reliably estimated from situations describing the confidence with which results have been obtained. A set of situations will be introduced, describing semantic confidence based on the agreement of semantic interpretations proposed by different classification methods: decision-tree based classifiers (Semantic Classification Trees, SCT [2]) and large-margin classifiers using boosting (BoosTexter [3]) and Support Vector Machines (SVM-Torch, [4]).

In [5] a method has been proposed for computing the probability that an interpretation Γ obtained from each sentence in a n-best list of conceptual interpretations is correct given the rank of the sentence and a set of acoustic, linguistic and semantic confidence features. The computation of this probability is often highly imprecise. A better strategy for selecting an interpretation is introduced in this paper by taking into account the results of redundant interpretation processes using different classifiers, using confidence measures taken mostly just on the words that support the generation of semantic constituent hypotheses. The interpretation strategy considers, with the highest priority, the validation of the interpretation arising from the most likely sequence of words. If the probability that this interpretation is not high, then possible corrections of it are considered using the interpretations emerging from the other sequences in the N-best lists.

2. KNOWLEDGE REPRESENTATION AND USE

The type of applications considered in this paper are telephone services in which a user performs requests to a system. Following definitions and notations in [6], an example of user request is represented by an instance of a speech act of the type:

```
REQUEST(user,service,
   INFORMREF(user,service,x,
       (SATISFIES(Restaurant(x)),Path(PTH))))
```

where `REQUEST` indicates an illocutionary speech act, user and service are the conversant entities, `INFORMREF` is a propositional content expressing the constraints of the request. In this example, the constraints apply to the request for instances `x` of the semantic structure Restaurant in an area described by an instance `PTH` of the semantic structure `Path`. Other types of speech acts are `INFORM`, `QUESTION`.

In order for the service dialogue strategy to provide a suitable action, the speech understanding subsystem (SUS) has to hypothesize the type of dialogue act, and the propositional content. For this particular type of application, the conversant entities are constant and it may not be necessary to identify the user, although having some information about him/her might be useful.

The propositional content is obtained by the interpretation strategy of the SUS which generates the instance `PTH` and the fact that there is a request about a restaurant whose answer should satisfy `PTH`. Generation of semantic interpretation is a process of evidential reasoning in which composition and inference are based on semantic knowledge expressed by an appropriate formalism and on probabilities for computing the likelihood of a result given the imprecision of hypotheses and knowledge.

Most of the approaches proposed so far for Spoken Language Understanding (SLU) integrate semantic knowledge into a context-free semantic grammar and propose different algorithms for computing the probability $P(\Gamma, W)$ of a conceptual structure Γ and a sequence of words W [7, 8, 9]. Context-free semantic grammars

This research is supported by France Telecom's R&D under the contract 021B178

have nonterminal symbols which represent semantic structures and can be rewritten into non-overlapping sequences of words.

Interesting books [10, 11] describe various types of semantic knowledge and their use. A common aspect of many of them is that it is possible to represent complex relational structures with non-probabilistic schemes that are more effective than context-free grammars. For example, in K-LONE [12] concept descriptions account for the internal structure with Role/Filler Descriptions, called Roles and for a global Structural Description (SD). Roles have substructures with constraints specifying types and quantities of fillers. SDs are logical expressions indicating how role fillers interact. Role descriptions contain value restrictions. Epistemological relations are defined for composing conceptual structures. They may connect formal objects of the same type and account for inheritance. It is important to point out that semantic knowledge is, in general, context-sensitive.

Instances containing token values are the basis for the inference performed by the dialogue strategy. The inference process uses ontologies and relations among semantic components and is based on algorithms for theorem proving or spreading activation on semantic networks. It is worth noticing that for representation formalisms like KL-ONE, inheritance can take place for concepts, roles as well as structural descriptions. Automatic reasoning is performed with these schemes using methods such as theorem proving, rule chaining or spreading activation. Evidential reasoning with these methods appear to be complex and often not feasible in a rigorous way.

3. SPEECH UNDERSTANDING STRATEGY

Following the semantic models presented in the previous section, the first step in the semantic interpretation process is to detect and extract the concepts (or speech acts) needed in order to fill the semantic structure considered. These concepts represent speech acts like:
request_for_help, *ask_for_repeat*, *ask_for_information*, *access_to_a_list_item*, *confirmation*, *negation*,...
and entities that are associated with specific values:
restaurant_location, *specialties*, *price*, *time*,....

In order to detect such concepts from the ASR output, we use two kinds of models:

- text classifiers that are trained to label a string of words with one or several concept tags;
- regular grammars coded as Finite State Machines (FSMs), only on the concepts representing the entities with values. There is one FSM for each kind of entity (*price, specialties,...*) that models all the different values accepted by the entity.

The general Speech Understanding strategy, presented in figure 1, can be described as follows:

1. the ASR module outputs a word graph;
2. a *Structured N-best* list of hypotheses [1] is extracted from the word graph, containing the n-best conceptual interpretations contained in the word graph as well as the n-best word hypothesis for each interpretation
3. the best hypothesis of the first interpretation is selected as the reference hypothesis $W_{1,1}$;
4. each classifier processes $W_{1,1}$ in order to output a string of concept tags;
5. a decision rule checks the agreement between the classifiers on the concept strings outputs;
6. if all classifiers agree, $W_{1,1}$ with its concept string is transferred to the dialogue manager in order to build the semantic representation of the utterance;
7. if no total agreement is reached, an error correction module looks for the most reliable correction of $W_{1,1}$ within the *Structured N-best* list. A score is given to the chosen correction, the utterance can be rejected according to a threshold on this score.

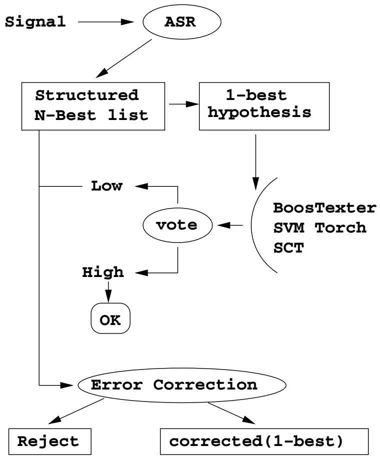

Fig. 1. Architecture of the error correction strategy

Before presenting the probabilistic model behind our error correction module, we describe first in the next section the classification tools used in the decision process.

4. SPEECH UNDERSTANDING WITH TEXT CLASSIFIERS

Several studies have shown that text classification tools (like Support Vector Machines or Boosting algorithms) can be an efficient way of labeling an utterance transcription with a semantic label such as a call-type [13] in a Spoken Dialogue context. This approach has two main advantages: firstly, the amount of human supervision is limited as no keywords or grammars have to be defined in order to characterize a concept. The only manual data needed is a training corpus containing, for each utterance, the string of concepts occurring in it. Secondly, classifiers are more robust to the noise generated by ASR errors and spontaneous speech effects. Indeed, they can be trained directly from ASR output and therefore model this noise.

In the proposed strategy we use three kinds of classifiers: a decision-tree based classifier (Semantic Classification Trees or SCT [2]) and two large-margin classifiers, one based on Support Vector Machine (SVM-Torch, [4]) and one implementing a boosting method of weak classifiers (BoosTexter [3]). Because these tools are based on different classification algorithms with different input formats (bag-of-words or word-strings for example), they don't always use the same information in order to characterize a label. Therefore, using them simultaneously and defining a voting decision rule, can increase the classification robustness.

5. ERROR CORRECTION PROBABILITIES

We present here the probabilistic model behind the error correction strategy presented in section 1.

Let Γ_1 be the most likely interpretation from the Structured N-best list and $W_{1,1}$ be the most likely sequence of words correspondig to Γ_1. Better semantic hypotheses have to be searched for in the other word sequence candidates only when there is a high probability that this search results in a better interpretation. It is thus important to find conditions for evaluating the probability that Γ_1 is correct. Furthermore, even when Γ_1 is incorrect, it is likely that it is partially correct and a better interpretation has to be found as a correction of the errors in Γ_1.

Let us consider only modifications $\Gamma_c = m_c(\Gamma_1)$. Γ_c is an alternative in the Structured N-best list to Γ_1. The differences between these two interpretations (called modifications m_c) are insertions, substitutions or deletions of concepts between Γ_1 and Γ_c. Let $W_{c,1}$ be the best sequence of words expressing Γ_c.

In principle, semantic interpretations Γ should be scored with the following probability, with Y being the speech signal and W_g the word graph:

$$P(\Gamma \mid Y) = P(\Gamma \mid W_g)P(W_g \mid Y) \tag{1}$$

The following approximation is proposed to find the best interpretation:

$$\begin{aligned} P(\Gamma_c \mid W_g) \approx \quad & P\{m_c \mid \Gamma_1, W_{1,1}, S_1, F(W_{1,1}, W_{c,1})\} \\ P(\Gamma_1 \mid W_g) \approx \quad & \qquad\qquad P\{\epsilon \mid S_1, F(W_{1,1})\} \end{aligned} \tag{2}$$

S_1 is a confidence predicate about Γ_1. It is defined by agreement rules between the classifiers as presented in figure 1. Two levels of confidence are defined: high confidence (*High*) if all classifiers agree and low confidence (*Low*) if they don't. W_g is represented by $W_{1,1}$ and a set of measures or confidence predicates F taken on $W_{1,1}$ and $W_{c,1}$. The confidence measures F include acoustic and linguistic features as described in [5]. ϵ means that no correction is done on Γ_1.

Error correction I_q can be seen as special type of inference in which a new interpretation $T_q(\Gamma_q)$ is obtained from the interpretation Γ_q when some premise PR_q is true. The general form for the q-th type of error correction is:

$I_q : \Gamma_q \wedge PR_q \rightarrow T_q(\Gamma_q)$

where PR_q are expressions conditioning the application of a correction.

If corrections apply only to Γ_1, then: $\Gamma_q = \Gamma_1$ and $I_q = m_c(\Gamma_1)$. On the other hand, $I_q = \epsilon(\Gamma_1)$ means that Γ_1 is accepted.

Decision on applying corrections depends on the following probability:

$$P(T_q \mid PR_q) = \frac{P(PR_q \mid T_q)P(T_q)}{P(PR_q \mid T_q)P(T_q) + P(PR_q \mid \neg T_q)P(\neg T_q)} \tag{3}$$

where $\neg$ means negation.

All these probabilities are learned on a development corpus in the following way: firstly, a Structured N-best list is estimated on each utterance of the corpus. Secondly, each hypothesis of these lists is labeled with a tag indicating that the interpretation attached to the hypothesis is correct or not and with a set of confidence measures, as presented in [5]. Finally a decision-tree training is performed on these labeled hypotheses in order to separate the correct interpretations from the wrong ones.

Decision is based on the value of $P\{(T_q = \epsilon(\Gamma_1)|PR_q\}$. Each path in the decision tree corresponds to a premise PR_q. Each leaf has also associated counts from which the probability $P\{(T_q = \epsilon(\Gamma_1)|PR_q\}$ can be estimated. If the probability is greater than a threshold, then Γ_1 is considered to be correct. If the probability that Γ_1 is correct is too low, then we look for the T_q which maximizes the probability $P(T_q \mid PR_q)$. According to the probability obtained we can decide to either accept the correction or reject the utterance.

6. EXPERIMENTAL RESULTS

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. The language model used is made of 44K words. The interpretation strategy was inferred using a development corpus containing 2.1k utterances. Performance was evaluated on a test corpus containing 1.7k utterances. The Word Error Rate (WER) on the development and test corpora, considering the best word sequence obtained only with the generic LM, were 25.83% and 26.98% respectively. The measure considered here is the Understanding Error Rate (*UER*) that is related to the tags and the values of the concepts detected. 15 concept tags are used in these experiments. In order to evaluate the performance of the classifiers alone, a Classification Error Rate (*CER*) has also been evaluated that just takes into account the concept tags but not their values.

test	*ref*	*asr*	*coverage*
best class.	3.2%	9.6%	100%
agreement	-	5.8%	72%

Table 1. Classification Error Rate (CER) on the concept tags obtained by the best text classifier and by the agreement of all the classifiers. *ref* is the reference transcription of the utterances, *asr* is the output of the ASR module, *coverage* is the percentage of the corpus on which the classification results are given

Table 1 shows the CER results, on the test corpus, by means of the best text classifier. As we can see the performance is excellent on the reference transcriptions (around 97% of the tags are correct). This indicates that text classifiers are well suited for this understanding task. The drop in performance due to the noise generated by the ASR errors is also a strong indicator that a correction strategy needs to be added in order to recover some of these errors. If we look at the classification rate of the utterances that received a *High* confidence label (agreement of all the classifiers), we obtain a CER of 5.8% and the percentage of occurrences labeled with *High* confidence is 72%. This indicates that the agreement rule among the classifiers is a powerful confidence measure.

Table 2 presents the UER results obtained at different steps in our strategy: the column **High** shows the performance of the agreement decision rule (among the different classifiers). 72% of the test corpus utterances have been labeled with a high confidence. In this case we keep the first hypothesis extracted from the Structured N-best list (named Γ_1, W_1) and the UER on these utterances is only 6.2% (5% on the development corpus). On the other utterances, labeled with a **Low** confidence, the UER without error correction is 23.6% (27.2% on the development corpus). This result clearly

Development corpus					
Conf	**High**	**Low**		**All**	
Γ, W	Γ_1, W_1	Γ_1, W_1	Γ_c, W_c	Γ_1, W_1	Γ_c, W_c
UER	5%	27.2%	17.8%	10.9%	8.4%
size	74.3%	25.7%		100%	

Test corpus					
Conf	**High**	**Low**		**All**	
Γ, W	Γ_1, W_1	Γ_1, W_1	Γ_c, W_c	Γ_1, W_1	Γ_c, W_c
UER	6.2%	23.6%	18.0%	11.3%	9.7%
size	72%	28%		100%	

Table 2. Understanding Error Rate (UER) according to the confidence given by the classifiers (**High** or **Low**) and the strategy used: (Γ_1, W_1) means that the 1-best hypothesis is chosen, (Γ_c, W_c) means that the error correction strategy is applied (only on the **Low** confidence utterances). **size** indicates the % of the corpus labeled with **High** or **Low** confidence. **All** reports the overall error correcting performance

shows that agreement between classifiers is a very good predicate for estimating the confidence of an hypothesis.

By applying the error correction strategy presented in this paper, the UER drops to 18%, which is a relative improvement of 23.7% (34% of relative improvement on the development corpus). The overall drop in understanding errors is about 10% on the whole test corpus and 23% on the whole development corpus.

Furthermore we can use $P(I_q \mid PR_q)$ presented in section 5 in order to implement a rejection strategy: by fixing a threshold δ on this probability, we can reject utterances that are labeled with low confidence, when all possible corrections I_q have a probability lower than δ.

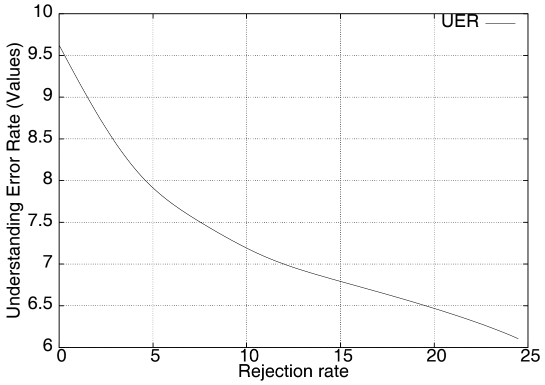

Fig. 2. Understanding Error Rate vs. Utterance Rejection Rate with the error correction strategy on the test corpus

This rejection strategy is evaluated on figure 2 on the test corpus. As we can see, by fixing an operating point at 10% utterance rejection rate, the UER drops from 9.7% to 7.2% which is relative improvement of about 25%.

7. CONCLUSION

This paper presents an error correction strategy applied to semantic interpretation of utterances in a Spoken Dialogue context. Several classifiers are used in order to give confidence labels to interpretations output by the Spoken Language Understanding module. If the confidence is low, a decision-tree specifically trained in order to estimate the best correction possible according to a set of acoustic, linguistic and semantic confidence measures is applied to the 1-best concept and word hypothesis. Results obtained on a dialogue corpus provided by France Telecom R&D and collected for a tourism telephone service show significant reduction in understanding error rate. Moreover, the error correction probabilities prove to be efficient new confidence measures that can be used in an utterance rejection decision rule.

8. REFERENCES

[1] Christian Raymond, Yannick Estève, Frédéric Béchet, Renato De Mori, and Géraldine Damnati, “Belief confirmation in spoken dialogue systems using confidence measures,” in *ASRU’03*, St. Thomas, US-Virgin Islands, 2003.

[2] R. Kuhn and R. De Mori, “The application of semantic classification trees to natural language understanding,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17, no. 449-460, 1995.

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

[4] R. Collobert, S. Bengio, and J. Mariéthoz, “Torch: a modular machine learning software library,” in *Technical Report IDIAP-RR02-46, IDIAP*, 2002.

[5] Christian Raymond, Frédéric Bechet, Renato De Mori, Géraldine Damnati, and Yannick Esteve, “Automatic learning of interpretation strategies for spoken dialogue systems,” in *Proc. IEEE ICASSP’04*, Montreal, Canada, 2004.

[6] J. Allen, “Natural language understanding,” *Benjamin/ Cummings publ. co.*, 1988.

[7] E. Levin E. and R. Pieraccini R., “Concept-based spontaneous speech understanding system,” in *Proceeding Eurospeech-95*, Madrid, Spain, 1995, pp. 555–558.

[8] Y. He and S. Young, “Hidden vector state model for hierarchical semantic parsing,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, 2003, pp. 268–271.

[9] Y. Gao R. Sarikaya and M. Picheny M., “A comparison of rule–based and statistical methods for semantic language modeling and confidence measurement,” in *Proc HLT-NAACL Conference*, Boston, USA, 2004, pp. 65–68.

[10] R. Jackendoff, “Semantic structures,” *The MIT Press, Cambridge Mass.*, 1990.

[11] H. J. Levesque and R.J. Brachman, “A fundamental tradeoff in knowledge representation and reasoning,” *In Readings in Knowledge representation, Morgan Kaufmann publ*, pp. 42–70, 1985.

[12] R. Brachman and J. Schmolze, “An overview of the KL-ONE knowledge representation system,” *Cognitive Science*, vol. 9(2), pp. 171–216, 1985.

[13] Patrick Haffner, Gokhan Tur, and Jerry Wright, “Optimizing SVMs for complex call classification,” in *Proc. IEEE ICASSP’03*, Hong-Kong, 2003.