Conceptual decoding from word lattices: application to the spoken dialogue corpus MEDIA

Christophe Servan, Christian Raymond
Frédéric Béchet, Pascal Nocéra

LIA - University of Avignon, BP1228 84911 Avignon cedex 09 France
{*christophe.servan,christian.raymond,frederic.bechet,pascal.nocera*}*@univ-avignon.fr*

Abstract

Within the framework of the French evaluation program MEDIA on spoken dialogue systems, this paper presents the methods proposed at the LIA for the robust extraction of basic conceptual constituents (or concepts) from an audio message. The conceptual decoding model proposed follows a stochastic paradigm and is directly integrated into the Automatic Speech Recognition (ASR) process. This approach allows us to keep the probabilistic search space on sequences of words produced by the ASR module and to project it to a probabilistic search space of sequences of concepts. This paper presents the first ASR results on the French spoken dialogue corpus MEDIA, available through ELDA. The experiments made on this corpus show that the performance reached by our approach is better than the traditional sequential approach that looks first for the best sequence of words before looking for the best sequence of concepts.

Index Terms: Automatic Speech Recognition, Spoken Dialogue, Spoken Language Understanding.

1. Introduction

within the framework of Spoken Dialogue Systems, the French evaluation program MEDIA [1] is focused on the evaluation of conceptual decoding methods that translate a string of words into a string of *semantic concepts* with values. This evaluation has been made on manual transcriptions of spoken dialogues obtained thanks to a Wizard of Oz paradigm. The task targeted is touristic information and hotel booking application. In addition to this program this paper presents the first evaluation done on the audio files of the MEDIA corpus. By using the Automatic Speech Recognition (ASR) system SPEERAL [4] and the Spoken Language Understanding (SLU) module developed at the LIA [5] we show how an *integrated* approach, looking at the same time for both the best word string and the best conceptual interpretation of a message can outperform the traditional sequential approach looking first for the best word string, then for the best interpretation.

This paper is structured as follows: section 2 introduces the MEDIA corpus; the ASR model used for processing the audio corpus are presented in section 3; section 4 gives an overview of the SLU approach developed at the LIA; finally an evaluation of our models is given in section 5 on the MEDIA corpus. The two issues that are going to be addressed in the discussion of the experimental results are:

- what is the impact of the Word Error Rate (WER) on the performance of the SLU module?
- what are the gains, if any, in terms of both WER and Concept Error Rate (CER) achieved by the *integrated* approach proposed in this paper compared to the sequential approach?

2. The MEDIA corpus

The evaluation program MEDIA [1] aims to evaluate different conceptual decoding systems within the framework of a spoken dialogue system dedicated to provide touristic information and hotel booking service. A 1250 dialogue corpus has been recorded by ELDA following a Wizard of Oz protocol: 250 speakers have followed each 5 hotel reservation scenarios. This corpus has been manually transcribed, then conceptually annotated according to a semantic representation defined within the project. This representation is based on the definition of *concepts* that can be associated to 3 kinds of information:

- first a concept is defined by a label and a value; for example to the concept *date* can be associated the value *2006/04/02*;
- then a *specifier* can be attached to a concept in order to link the concept together in order to go from a flat concept-value representation to a hierarchical one; for example, to the concept *date* can be associated the specifiers *reservation* and *begin* to specify that this date is the beginning date of an hotel reservation;
- finally a *modal* information is attached to each concept (positive, affirmative, interrogative or optional).

n	W^{c_n}	c_n	*value*
0	uh	null	
1	yes	answer	yes
2	the	RefLink	singular
3	hotel	BDObject	hotel
4	which	null	
5	price	object	payment-amount
6	is below	comparative-payment	below
7	hundred and ten	payment-amount-int	110
8	euros	payment-currency	euro

Table 1: Example of message with concept+value information

Table 1 shows an example of message from the MEDIA corpus with the concept-value information only. The first column contains the segment identifier in the message, the second column shows the string of words W^{c_n} supporting the concept c_n of the third column. In the fourth column is displayed the value of the concept c_n in the string W^{c_n}.

The semantic dictionary MEDIA contains 83 concept labels, 19 specifiers and 4 modal information. In this study we will focus on the concept-value extraction only. No specifiers or modal information is considered, as these labels are given by another module of our SLU system. So the tagset considered is made of 83 labels. Let's note that this is quite a large tagset compared to previous spoken dialogue corpora like the ATIS or ARISE ones.

The MEDIA corpus is split into 6 parts. The first 4 ones (720 dialogues, 12K messages) are used for training the models, the 5th one (79 dialogues, 1.3K message) is used as a development corpus for tuning the models are the 6th one is the test corpus made of 200 dialogues.

3. ASR models used for processing the MEDIA corpus

3.1. Model training

The SPEERAL ASR system [4] has been used for processing the messages of the MEDIA corpus. These messages have been recorded in *real* conditions (i.e. the same ones as would find a deployed system), with a large variety of male and female speakers (250), no constraints on the caller (cell phone, surrounding noise), and therefore the acoustic quality of the messages is variable.

The acoustic models used have been trained on the ESTER corpus [2] and adapted to the MEDIA training corpus through a MAP adaptation process.

The Language Model (LM) used is a 3-gram LM trained on the manual transcriptions of the MEDIA training corpus. This corpus is made of 226K words. The lexicon extracted from it contains 2028 words phonetically transcribed by means of the grapheme-to-phoneme transcription tool LIA_PHON[1]. On the MEDIA test corpus the Out-Of-Vocabulary rate with this lexicon is 1.6%, the perplexity of the 3-gram LM is 26.5.

The Word Error Rate (WER) obtained on the test corpus with such models is 33.5%.

3.2. Word lattices

The SLU approach proposed in this paper takes as input word lattices. These lattices, produced by the ASR decoder SPEERAL, are represented by Finite State Machines (FSM). All the further operations on these FSMs are performed thanks to the toolkit *AT&T FSM/GRM Library* [3]. Because one of the goals of this work is to study the correlation between the error rate on the words and the one on the concepts, we used these lattices in order to simulate different word recognition performance. The following method is used:

- first the word lattices generated by SPEERAL are produced for each message of the test corpus; by taking the best word string in these lattices an average WER of 33.5% is achieved;
- then a 3-gram language model is trained on the manual transcription of the test corpus;
- this new LM, represented as an FSM, is composed with a fusion factor to the word lattices already produced in order to rescore each word string;
- by tuning this fusion factor one can control the WER of the best word string of these lattices.

With this method we obtain 4 different decoding of the MEDIA test corpus with 4 different fusion factors (0.0 0.5 0.8 and 1.0). Four sets of word lattices are obtained: $G_{0.0}$, $G_{0.5}$, $G_{0.8}$ and $G_{1.0}$. The lattices $G_{0.0}$ correspond to the *baseline* decoding models where no data from the test corpus was included in the models. The WER of the best word strings of these lattices are:

Lattices	$G_{0.0}$	$G_{0.5}$	$G_{0.8}$	$G_{1.0}$
WER	33.5	29.8	27.8	23.9

Even if adding test data in the models brings a bias in the evaluation method, we believe that this bias is reduced by introducing the test data only in the second phase of the ASR decoding: the errors and acoustic confusions due to the baseline model are still there. However the lattices $G_{0,0}$ are the most realistic ones, because they are produced without any introduction of test data. The other lattices are only useful in order to see the correlation between the transcription and the understanding error rate.

4. Interpretation strategy

Let's note C the conceptual interpretation of a message. C is a sequence of basic concepts, like those defined in the MEDIA semantic model and shown in the table1. Conceptual decoding is the process which consists in looking for the best concept string $\hat{C} = c_1, c_2, \ldots, c_k$ and their corresponding values from a string of acoustic observations O. If the value for each concept c_i can be extracted directly from the string of word W^{c_i} supporting c_i, then conceptual decoding leads to find $\hat{W}, \hat{C}$ that are the best sequence of concepts and the best sequence of words from a string of acoustic observations O.

As presented in [7], this process can be done either sequentially in a *two-pass* approach where ASR models look first for the best word string $\hat{W}$ before processing this word string in order to find $\hat{C}$; or with a single pass approach like in the following formula:

$$P(\hat{C}, \hat{W}|O) \approx \max_{C,W} P(O|W)P(W,C) \qquad (1)$$

In this case the understanding model is used as a language model for the ASR process. Several studies have compared the two approaches [8, 6] and found that if the WER increases with the integrated single-pass method, the understanding measures are improved. Compared to these previous studies the task targeted in this work is more complex (noisy telephone speech, large conceptual lexicon); this study aims to prove that a single-pass conceptual decoding method can be robust enough to deal with such complex data.

Because of robustness issues, the approach proposed here is not a *pure* single-pass approach: a first ASR activity generates a word lattice thanks to acoustic models and a 3-gram language model. Then this word lattice is transformed into an entity lattice coded as a word-to-concept transducer. Finally an HMM-based conceptual tagger finds:

$$P(\hat{C}, \hat{W}) = \max_{C,W} P(W,C) \qquad (2)$$

in the entity lattice. In the experiment section we compare this approach, called the *integrated* approach, to the traditional sequential 2-pass approach that consists in keeping only the best word string in the word lattice produced by the ASR models.

The transformation process of a word lattice into an entity lattice is described in [5] and briefly presented in the next section.

[1] LIA_PHON *http://www.lia.univ-avignon.fr/chercheurs/bechet/*

4.1. Entity lattice

In the Spoken Language Understanding module developed at the LIA, interpretation starts with a translation process in which stochastic Language Models are implemented by Finite State Machines (FSM) which output labels for semantic constituents. These semantic constituents are called *concept tags* and are noted c_i. They correspond to the 83 concept tags defined in the MEDIA ontology. To each concept tag c_i is attached the word string W^{c_i} supporting the concept and from which the concept value (e.g. date, proper name or numerical information) can be extracted. 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.

They can be manually written for domain-independent conceptual constituents (e.g. dates or amounts), or data-induced for the concepts specific to the MEDIA corpus. All these transducers are grouped together into a single transducer, called *Concept FSM*, which is the union of all of them. During the decoding of a message, a first ASR module generates a word graph (G_W). G_W is composed with the transducers *Concept FSM*; 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).

4.2. Conceptual tagger

In order to find the best sequence of concept tags $t_1, t_2, \ldots, t_n$ as well as the best sequence of words $w_1, w_2, \ldots, w_n$, an HMM tagger, also encoded as an FSM is trained on the MEDIA training corpus. This is a 3-gram HMM tagger:

$$\hat{C}, \hat{W} \sim \underset{C,W}{\mathit{argmax}} \prod_{i=1}^{n} P(w_i, c_i | w_{i-1}, c_{i-1}, w_{i-2}, c_{i-2})$$

Special labels are attached to each word: B for beginning of a concept entity, I for inside a concept and O for outside a concept. In order to increase the generalization power of the model, some categories of words are replaced by labels (e.g. numbers, month names, ...). During the decoding process of a message, this tagger is composed with the transducer T_{WC}.

4.3. Structured N-best list

The result of the understanding process is a *Structured N-Best* list of interpretations that can be seen as an abstraction of all the possible interpretations of an utterance. To produce this list a first n-best list of hypotheses is obtained on the output symbols of the transducer T_{WC}. This is a list of concept strings, with no values attached. Then, for the first m concept string hypotheses, an n-best list of values is obtained by considering the input symbols of T_{WC}. These values are obtained through another transducer called T_{value}. This strategy is resumed in figure 1.

5. Experiment

The experiment reported here are obtained on the test section of the MEDIA corpus, containing 200 dialogues, the concept lexicon used contains the 83 basic concepts of the MEDIA semantic model. Performance is reported according to the *Concept Error Rate* (CER). A concept is considered as correct only if both its tag and its value are correct according to the reference.

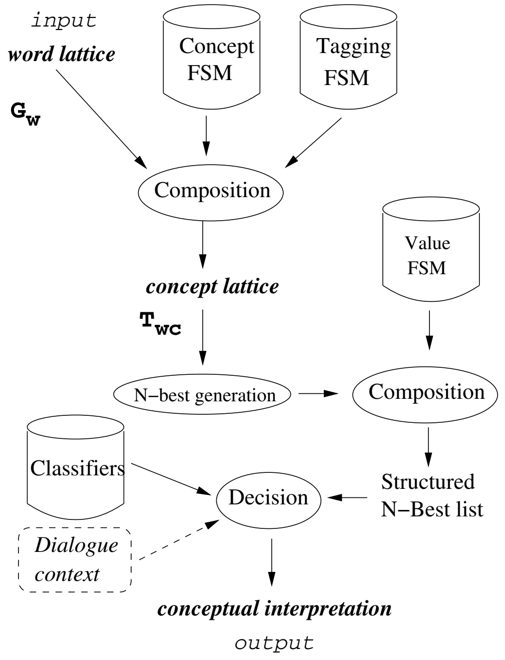

Figure 1: The LIA Spoken Language Understanding strategy

Table 2 compare the *integrated approach* to the *sequential approach* where only the 1-best word string in the word lattice is processed by the SLU module.

input	$G_{0.0}$		$G_{0.5}$		$G_{0.8}$		$G_{1.0}$	
meth.	seq	int	seq	int	seq	int	seq	int
WER	33.5	33.4	29.8	29.7	27.8	28.8	23.9	26.6
CER	42.6	38.9	40.8	38.5	39.1	37.0	37.5	35.5

Table 2: WER and CER for different word lattices with the sequential approach (seq) and the integrated one (int)

As we can see the understanding performance is improved with the integrated approach with a CER going from 42.6 down to 33.5 while the WER remained constant. With the other lattices, the integrated approach always brings a gain in understanding even if the WER is increased. Figure 2 shows the correlation between the WER obtained on the 1-best string of the different lattices and the CER obtain for both approaches.

In figure 3 3 different n-best list of hypotheses are compared through their *Oracle* CER for the lattices $G_{0.0}$. The *Oracle* CER is the minimal error rate that would achieve a perfect decision rule choosing a hypothesis among the n-best list. These 3 n-best list are:

- *N-Best 1*: this list is produced with the sequential approach by enumerating the best sequence of concepts $\hat{C}$ on the 1-best word string of $G_{0.0}$.
- *N-Best 2*: this list is obtained directly on T_{WC}, once the conceptual tagger has been applied, by enumerating the best paths in this transducer.
- *N-Best 2 struct.*: this corresponds to the structured n-best list [5] obtained with the integrated approach.

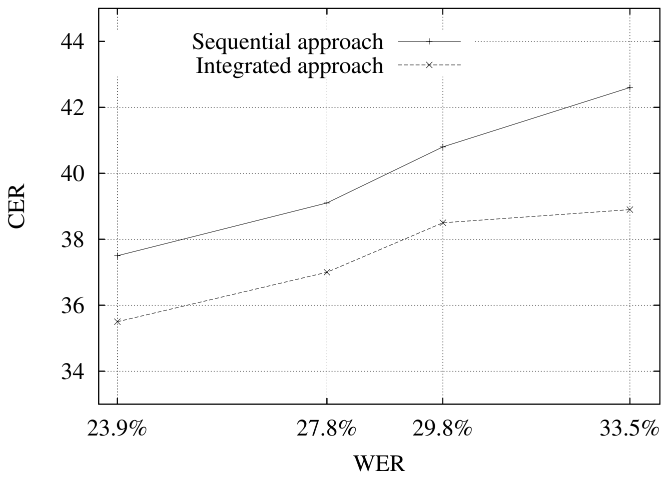

Figure 2: CER as a function of the WER for both sequential and integrated approaches

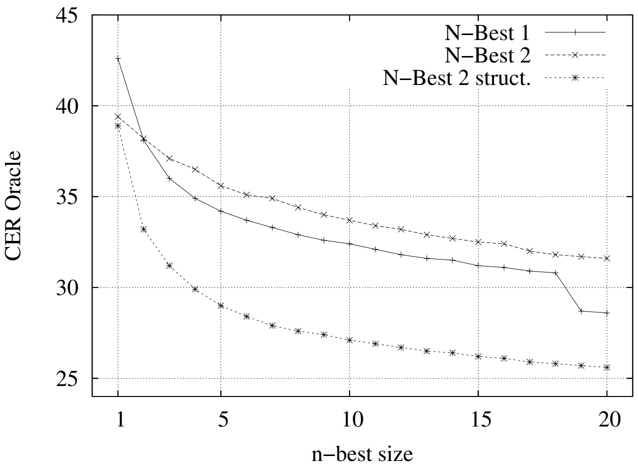

Figure 3: Oracle CER for lattices $G_{0.0}$ with the sequential approach (NBest 1), the integrated approach with standard n-best list (N-Best 2) and the structured N-Best list (N-Best 2 Struct.)

As we can see the Structured n-best list outperforms the other n-best lists: by keeping only the best 3 hypotheses, the same Oracle CER is obtained than all the n-best lists of the other methods.

6. Conclusion

We have shown in this study how a robust understanding process can be integrated into an ASR system, without using the usual suboptimal sequential approach that looks first for the best string of words before looking for the best string of concepts. The conceptual decoding model proposed follows a stochastic paradigm and is directly integrated into the ASR process. This approach allows us to keep the probabilistic search space on sequences of words produced by the ASR module and to project it to a probabilistic search space of sequences of concepts. The results obtained are the first ASR results on the French spoken dialogue corpus MEDIA, available through ELDA. The experiments performed on this corpus have shown that the performance reached by our approach is better than the traditional sequential approach.

Extracting a structured n-best list of hypotheses has also proven to be a very efficient way of reducing the amount of hypotheses that can be sent to the dialogue manager without reducing the Oracle performance.

7. References

[1] Helene Bonneau-Maynard, Sophie Rosset, Christelle Ayache, Anne Kuhn, and Djamel Mostefa. Semantic annotation of the french media dialog corpus. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Lisboa, Portugal, 2005.

[2] G. Gravier, J.F. Bonastre, E. Geoffrois, S. Galliano, K. McTait, and K. Choukri. ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français. In *Proc. Journées d'Etude sur la Parole (JEP)*, 2004.

[3] Mehryar Mohri, Fernando Pereira, and Michael Riley. Weighted finite-state transducers in speech recognition. *Computer, Speech and Language*, 16(1):69–88, 2002.

[4] P. Nocera, G. Linares, and D. Massonie. Principes et performances du décodeur parole continue Speeral. In *Proc. Journées d'Etude sur la Parole (JEP)*, 2002.

[5] Christian Raymond, Frdric Bchet, Renato De Mori, and Graldine Damnati. On the use of finite state transducers for semantic interpretation. *Speech Communication*, 48,3-4:288–304, 2006.

[6] G. Riccardi and A. L. Gorin. Stochastic language models for speech recognition and understanding. In *Proceedings of ICSLP*, 1998.

[7] Y. Wang, L. Deng, and A. Acero. Spoken language understanding. In *in IEEE Signal Processing Magazine. Volume: 22 Issue: 5*, pages 16–31, Sep. 2005.

[8] Y.-Y. Wang and A. Acero. Is word error rate a good indicator for spoken language understanding accuracy. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, page 577582, 2003.