SEMANTIC COMPOSITION PROCESS IN A SPEECH UNDERSTANDING SYSTEM

Frédéric Duvert, Marie-Jean Meurs, Christophe Servan, Frédéric Béchet, Fabrice Lefèvre, Renato De Mori

Laboratoire Informatique d'Avignon (LIA)
339, chemin des Meinajaries Agroparc BP 1228
84911 Avignon Cedex 9 France
{frederic.duvert, marie-jean.meurs, christophe.servan,
frederic.bechet, fabrice.lefevre, renato.demori}@ univ-avignon.fr

ABSTRACT

A knowledge representation formalism for SLU is introduced. It is used for incremental and partially automated annotation of the MEDIA corpus in terms of semantic structures. An automatic interpretation process is described for composing semantic structures from basic semantic constituents using patterns involving constituents and words. The process has procedures for obtaining semantic compositions and for generating frame hypotheses by inference. This process is evaluated on a dialogue corpus manually annotated at the word and semantic constituents levels.

Index Terms— Spoken language understanding, semantic structures, frames, conceptual decoding, semantic annotation, semantic inference.

1. INTRODUCTION

Semantics deals with the organization of meanings and the relations between signs or symbols and what they denote or mean. Spoken Language Understanding (SLU) is the interpretation of signs conveyed by a speech signal. Relations are represented by Knowledge Sources (KS) and applied by processes using control strategic knowledge. This task is difficult because meaning is mixed with other information like speaker identity and noise environment. Natural language sentences are often difficult to parse and spoken messages are often ungrammatical. The knowledge used is often imperfect and the transcription of user utterances in terms of word hypotheses is performed by an Automatic Speech Recognition (ASR) system which makes errors. In order to minimize the effects of imprecision, the interpretation has to be conceived as a decision process which can be conceptually decomposed into sub-tasks. It was observed [1] that an increase in precision may be achieved by computing a lattice of scored hypotheses of semantic constituents from a lattice of scored word hypotheses. Semantic constituents are further composed into semantic structures. Semantic constituent hypotheses are generated using stochastic finite state machines (FSM) along the line of research presented in [2, 3]. This paper describes a novel semantic composition and evaluation process which composes semantic constituents into semantic structures. Constituents are generated by a translation process from word lattices. Constituents and words have links to patterns. When patterns match with features based on constituent and word hypotheses, structure building procedures are executed. Confidence values based on probabilities are used for selecting hypotheses. The approach has been tested on a fairly complex French corpus called MEDIA, available through the ELDA corpus distribution agency.

This work is supported by the 6th Framework Research Programme of the European Union (EU), Project LUNA, IST contract no 33549. The authors would like to thank the EU for the financial support. For more information about the LUNA project, please visit the project home-page, www.ist-luna.eu.

2. THE MEDIA CORPUS AND THE GENERATION OF BASIC CONSTITUENT HYPOTHESES

2.1. Corpus description

The MEDIA corpus [4] has been recorded using a *Wizard of Oz* system simulating a telephone server for tourist information and hotel booking . Eight scenario categories were defined with different levels of complexity. The corpus accounts 1257 dialogs from 250 speakers and contains about 70 hours of dialogs. The training portion of the corpus is conceptually rich with more than 80 basic concepts manually transcribed and annotated. This *flat* semantic representation is enriched with labels that can be seen as traces of the underlying hierarchical representation. Hierarchical semantic representation is powerful as it allows to explicitly representing relationships between segments, possibly non-adjacent in the transcription of the query. On the other hand, a flat representation facilitates the manual annotation of the data. It has then been decided for the MEDIA annotation scheme to preserve the relationships, by defining a set of *specifiers* which are combined with the basic roles. There are 19 specifiers in the MEDIA semantic model.

n	W^{c_n}	c_n	*mode*	*specifier*	*value*
1	*I' m going to book*	command	+		reservation
2	*this hotel hotel Richard Lenoir*	hotel-name	+		richard lenoir
3	*six*	room-amount	+	reservation	6
4	*single rooms*	room-type	+		single
5	*for May thirty first*	date	+	reservation	31/05
6	*two days hum two nights*	night-amount	+	reservation	2

Table 1. Example (translated from French) of MEDIA semantic annotation on the message : *well hum I' m going to book this hotel hotel Richard Lenoir so six single rooms for May thirty first two days hum two nights*

An example of the MEDIA annotation on a message translated from French (*well hum I' m going to book this hotel hotel Richard Lenoir so six single rooms for May thirty first two days hum two nights*) is given in table 1. As we can see the specifier *reservation* is given to the concepts *command*, *room-amount*, *date* and *night-amount* as a hierarchical structure that would represent a reservation is triggered by the concept *command* and filled with the elements found in *room-amount*, *date* and *night-amount*. The combination of the specifiers and the attribute names allows recomposing a hierarchical representation of a query from its flat annotation, as it is going to be presented in this paper. This annotation provides labels comparable to semantic constituents hypothesized by a semantic shallow parser. The combinations of basic roles and the specifiers result in 1121 potential attributes. A total of 144 distinct attributes appears in the training corpus with about 2.2k different normalized values.

2.2. Conceptual decoding for generating basic constituents

The MEDIA corpus is annotated with basic semantic constituents but not with semantic structures. Basic semantic constituents are hypothesized and scored using an approach described in [1].

The conceptual decoding process is seen as a translation process in which stochastic Language Models are implemented by Finite State Machines (FSM) which output labels for semantic constituents. 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. At decoding time they are applied to the word graphs output by the ASR decoder by means of a composition operation. In order to find the best sequence of concept tags and words, an HMM tagger, also encoded as an FSM is used to rescore every path in the word/concept graph obtained. This HMM tagger is trained on the MEDIA training corpus. This approach is called an *integrated* decoding approach as the ASR and SLU processes are done together by looking at the same time for the best sequence of words and concepts. The result of the translation process is a *structured* n-best list of interpretations that can be seen as an abstraction of all the possible interpretations of an utterance.

2.3. Adding specifier labels to concept sequences

The conceptual interpretations from the n-best list produced don't contain any specifier labels. These specifiers are given in a second phase, on the hypotheses produced, by a tagging process based on *Conditional Random Fields* (CRF) [5]. CRF have been widely used for various word labelling tasks such as Part-Of-Speech tagging or Named Entity detection. CRF is a discriminant approach, it has proven to give better results on these tasks than generative HMM-based approaches. The main advantage of CRF is the ability to predict a word label according to a whole set of features related to the entire message, and not just the short history of the word to tag. This is very important for the task of adding specifiers to concepts as this information depends on features that can be far away from the concept to tag in the message.

The CRF specifier tagger is trained on the MEDIA corpus, each message is a sequence of features (words, attributes, values), labelled with a specifier label or the symbol *NULL*. At decoding time each word/concept sequence hypothesis of the structured n-best list is processed by the tagger in order to add these specifier labels. The CRF tool used is **CRF++** [1].

3. COMPOSING SEMANTIC RELATIONS INTO STRUCTURES

Semantic structures can be derived from semantic knowledge obtained with a semantic theory. Examples are semantic networks to represent entities and their relations [6], function/argument structures [7] and others. A convenient way for representing and reasoning about, semantic knowledge is to represent it as a set of *logic formulas* from which computational structures such as frames can be derived. A frame is a model for representing semantic entities and their properties. Frames should be able to represent types of conceptual structures as well as instances of them. Part of a frame is a data structure which describes the properties of a semantic structure, the constraints which should be respected by the values the property can assume, and procedures for obtaining property values from signs coded in the speech signal. In practice,

[1] http://crfpp.sourceforge.net/

properties are seen as slots to be filled by attached procedures with values called *slot fillers*. Slot filler can be the instance of another frame. This is represented by a pointer from the filler to the other frame. By filling slots, frame instances are generated. Acceptable frames for semantic representations in a domain can be characterized by a *frame grammar*.

4. PROGRESSIVE ANNOTATION OF THE CORPUS IN TERMS OF SEMANTIC STRUCTURES

A frame based knowledge source (KS) was manually composed to describe the semantic composition knowledge of the MEDIA application. Some frames describe generic knowledge like spatial relations, some others are application specific. These frames were defined according to the *Berkeley FrameNet* paradigm. Figure 1 shows an example of projection from FrameNet to the Media KS.

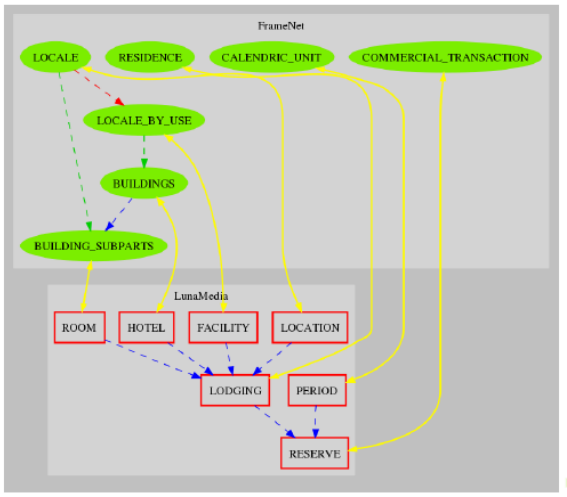

Fig. 1. Frame representation, projection from FrameNet to Media

This KS is composed of 21 basic frames with a total of 85 roles. The meaning representation language (MRL) contains conceptual constituents and semantic structure building procedures. These procedures are part of the semantics of the MRL. Semantic constituents and some words have links to patterns π_j. Patterns are made of constituent symbols, words and can include features extracted from the compounds of them. When a pattern matches with the incoming data, frame instantiations are created. Based on frame instances, inferences are performed. Different frames linked by relations may be instantiated by a single pattern.

An initial set of 463 turns of 15 dialogues was manually annotated with an average annotation time of 2 hours per dialogue. The FrameNet [8] annotation format was used. For example, the sentence "*I accept the reservation*" is annotated with three frames:

```
ACCEPT[(is_a:verb)(subject:person)(theme:reservation)]
PERSON [(is_a :human_being) (category : user) ..]
RESERVATION [(is_a : domain_object) ..]
```

Patterns were generalized by progressively annotating data with available knowledge, evaluating confidence of the results, manually annotating samples with low confidence and so on.

Attached procedures were integrated into an interpretation process to automatically provide frame annotations on the train corpus and instance hypotheses with the test corpus. The process is capable of performing inferences about frames whose instance is implied by other instantiated frames. Hundreds of rules generate instances from combinations of word patterns and semantic constituents and perform inferences on the results. There are 30 inference formulas used by the process. Annotations are described in XML documents containing additional information such as time references of words in patterns supporting hypotheses. A frame visualization tool, called FriZ, dedicated to process speech dialogues was developed to help manual annotation and verification of annotations obtained automatically.

At decoding time, once the n-best list of interpretations is obtained, with specifier labels on the concepts, as presented in section 2, each word/concept sequence is analyzed thanks to the logical rules developed on the MEDIA training corpus. These rules use the attributes, the values and the specifiers obtained in the first decoding phase in order to infer the frames. This operation could also use information about other speech events, related, for example to pitch and information stored in an agenda of hypotheses generated in previous dialogue turns. These sources of information are not taken into account in the work described in this paper.

5. EXPERIMENTAL RESULTS

Tests were performed on a corpus of 1249 dialog turns for a total of 2938 constituents. Table 2 shows the error rates obtained after the conceptual decoding phase. As we can see, for a Word Error Rate of 30.3%, the attribute error rate is about 25%. Each further information (specifiers and normalized values) add roughly an extra 6% to the error rates. We can see that the Oracle error rates obtained by manually selecting the best hypotheses in the n-best list of interpretations (with $n = 20$) are lower by an absolute 8% toward the 1-best error rates.

The frame annotation automatically obtained on the output of the decoding process has also been evaluated. Since we didn't have manual frame annotations for the test corpus, we used the manual annotations on words and concepts (with specifiers and values) available in MEDIA to obtain this frame reference, by applying the composition and inference knowledge described in the previous section. A random sampling on the turns of the test set was performed for manually verifying the accuracy of this automatic structure annotation with

tokens	Corr	Sub	Del	Ins	ER	Oracle ER
word	75.9	15.3	8.8	6.2	30.3	22.5
attribute	85.0	8.7	6.3	10.3	25.3	19.2
+ specif	78.6	15.2	6.2	10.2	31.6	23.4
+ value	72.5	21.4	6.1	10.1	37.6	25.2

Table 2. Error rate (ER) and Oracle ER on the n-best list of interpretations for words, tokens and tokens with specifier labels and values

two human experts. An F-measure of 0.90 (a precision of 0.96 and a recall of 0.85) was obtained on 100 turns comparing automatically generated frame annotations of exact transcriptions and manual annotations. This good accuracy validates the references obtained automatically.

This composition and inference knowledge was then applied to the n-best list of interpretations automatically obtained after the conceptual decoding process. The evaluation was done by estimating the precision, recall and F-measure on the detection of the correct frame type compared to the automatic frame reference described above. The Oracle F-measure is given on the n-best list in figure 2. As we can see an F-measure of 0.92 (a precision of 0.90 and a recall of 0.94) was obtained on the 1-best hypothesis for the 1249 dialog turns. These high precision results show that this high-level semantic annotation (the frame name) is robust to ASR errors, and the errors are occurring mostly on the frame elements. The next step of this work is to fully exploit the n-best list of interpretations in order to correct the erroneous frame elements thanks to the dialogue context.

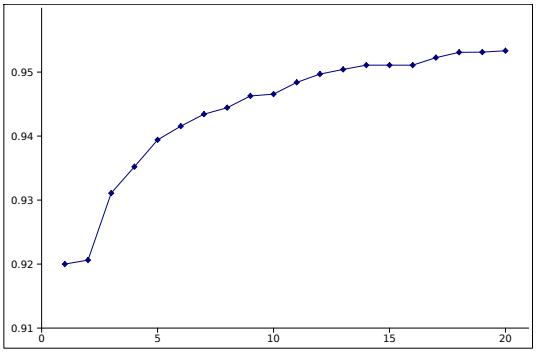

Fig. 2. Evolution, as function of N, of the Oracle F-measure for frame identification computed on the first N-best sequences of conceptual constituents extracted from the lattice of constituent hypotheses.

6. CONCLUSIONS

A knowledge representation formalism for SLU has been introduced. It has been used for incremental and partially automated annotation of the MEDIA corpus in terms of semantic structures. An automatic interpretation process has been introduced for composing semantic structures from basic semantic constituents using patterns involving constituents and words. The process has procedures for obtaining semantic compositions and for generating frame hypotheses by inference. Results in terms of F-measures are presented showing that the knowledge and the process have good capabilities for producing semantic structure hypotheses. Research will continue by using structural semantic knowledge for selecting possible constituents beyond the 1-best hypothesis in the whole lattice of concept hypotheses.

7. REFERENCES

[1] Christian Raymond, Frederic Bechet, Renato De Mori, and Geraldine Damnati, "On the use of finite state transducers for semantic interpretation," *Speech Communication*, vol. 48, no. 3-4, pp. 288–304, 2006.

[2] Giuseppe Riccardi and A. L. Gorin, "Stochastic language adaptation over time and state in natural spoken dialogue systems," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 3–10, 2000.

[3] Chai Wutiwiwatchai and Sadaoki Furui, "A multi-stage approach for thai spoken language understanding," *Speech Communication*, vol. 48, no. 3-4, pp. 305–320, 2006.

[4] 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.

[5] John Lafferty, Andrew McCallum, and Fernando Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th International Conf. on Machine Learning*. 2001, pp. 282–289, Morgan Kaufmann, San Francisco, CA.

[6] W.A. Woods, *What's in a Link: Foundations for Semantic Networks*, Bolt, Beranek and Newman, 1975.

[7] R. Jackendoff, "Semantic structures," *The MIT Press, Cambridge Mass.*, 1990.

[8] J.B. Lowe, C.F. Baker, and C.J. Fillmore, "A frame-semantic approach to semantic annotation," in *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, D.C., USA, April 1997.