Semantic composition process in a speech understanding system

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

SEMANTIC COMPOSITION PROCESS IN A SPEECH UNDERSTANDING SYSTEM

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

LIA - University of Avignon, 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 constituent 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, such as speaker identity or noise in the 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 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 [2]. 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 [3, 4].

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 the fairly complex French MEDIA corpus, 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. For more information about the LUNA project, please visit [1].

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

2.1. Corpus description

The MEDIA corpus [5] 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.

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 *com-*

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

Table 1. Example (translated from French) of MEDIA semantic annotation

mand, *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 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 following the approach described in [2].

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. 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 produced n-best list have no specifier labels. These specifiers are added in a second phase by a tagging process based on discriminant classifiers [6]. *Conditional Random Fields* (CRF) [7], retained in our study, have been widely used for various word labeling tasks such as Part-Of-Speech tagging or Named Entity detection. CRF is a discriminant approach, it has been shown 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++** [1] toolkit is used in this work.
CRFs capture long distance dependencies that support constituents of semantic structures without applying specific parsing rules.

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 [8] or function/argument structures [9]. A convenient way for representing and reasoning about semantic knowledge is to represent it as a set of *logic formulae* 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

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

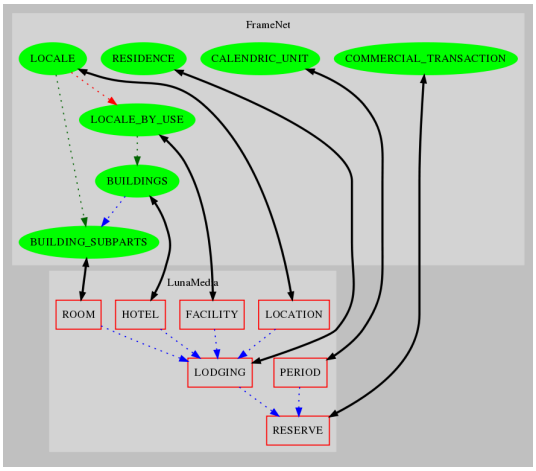

Fig. 1. Frame representation, projection from *FrameNet* to MEDIA

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, properties are seen as slots to be filled by attached procedures with values called *slot fillers*. A slot filler can be an 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 the semantic representation of a domain can be characterized by a *frame grammar*.

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

A frame based KS was manually composed to describe the semantic composition knowledge of the MEDIA domain. Some frames describe generic knowledge like spatial relations, some others are application specific. These frames were defined according to the *Berkeley FrameNet* paradigm adopted in [1]. Figure 1 shows an example of a semantic representation in the Media Corpus.

The MEDIA KS is composed of 21 basic frames with a total of 85 slots. 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 from 15 dialogues was manually annotated. The *FrameNet* [10] annotation format was used. A frame visualization tool, called FriZ, dedicated to process speech dialogues was developed to support manual annotation and verification of subsequent automatic annotations. The average manual annotation time per dialogue is around 2 hours. 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 and manually annotating samples with low confidence.

Attached procedures were integrated into an interpretation process to automatically provide frame annotations on the training 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 and semantic constituent patterns and perform inferences on the results. There are 30 inference formulae used by the process.

At decoding time, once the n-best list of interpretations is obtained with specifier labels 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 benefit from information related to other speech events, for example to the speaker pitch or to the hypotheses generated in the previous dialogue turns (stored in an agenda). These sources of information are not yet integrated 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 gives the error rates obtained after the conceptual decoding phase. 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. 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% than the 1-best error rates.

The frame hypotheses obtained on the output of the interpretation process has also been evaluated in view of. Since manual frame annotations were not available for the test corpus, the manual annotations of words and concepts were used to derive a reference frame annotation. After the composition and inference knowledge described in the previous section has been applied, a random sampling on the test user turns was performed by two human experts to manually assessing the accuracy of the automatic structure annotation. An F-measure of 0.90 (0.96 precision and 0.85 recall) was measured on 100 turns when comparing manual annotations

tokens	corr(%)	sub(%)	del(%)	ins(%)	ER(%)	Oracle ER(%)
word	75.9	15.3	8.8	6.2	30.3	22.5
concept	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 concepts and concepts with specifier labels and values

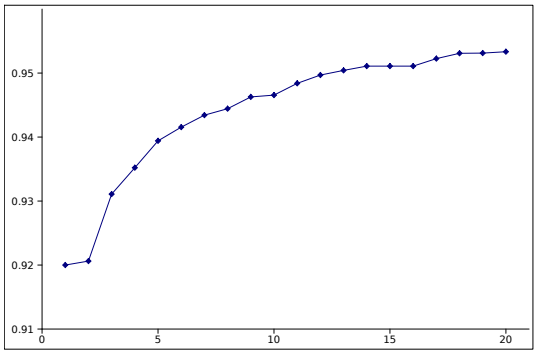

Fig. 2. Oracle F-measure for frame identification computed on the first n-best sequences of conceptual constituents extracted from the decoded lattice, as a function of n

and automatic frame annotations of exact transcriptions. This high accuracy allows to use the automatically-derived annotations as reference annotations.

The composition and inference knowledge was 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, using the automatic frame reference annotations described above. The Oracle F-measure is given on the n-best list in Figure 2. An F-measure of 0.92 (0.90 precision and 0.94 recall) was obtained on the 1-best hypothesis for the 1249 dialog turns. These results tend to show that the uppermost level of semantic annotation (frame identity) is pretty robust to ASR errors, the interpretation errors occurring mostly at the frame element level. The next step of the work will be to fully exploit the interpretation n-best list in order to correct the erroneous frame elements by consideration of the dialogue context.

6. CONCLUSION

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. Automatic annotations were evaluated and submitted to a human expert where confidence was low. 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. This research will be pursued by using structural semantic knowledge for selecting possible constituents beyond the 1-best hypothesis in the whole lattice of concept hypotheses.

7. REFERENCES

[1] "Project luna : www.ist-luna.eu," .

[2] 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.

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

[4] 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.

[5] Hélène Bonneau-Maynard, Sophie Rosset, Christelle Ayache, Anne Kuhn, and Djamel Mostefa, "Semantic annotation of the french media dialog corpus," in *Eurospeech*, Lisboa, Portugal, 2005.

[6] Fabrice Lefèvre, "Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation," in *ICASSP*, Hawaii, USA, 2007.

[7] John Lafferty, Andrew McCallum, and Fernando Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*. 2001, pp. 282–289, Morgan Kaufmann, San Francisco, CA.

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

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

[10] 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.