SPOKEN LANGUAGE INTERPRETATION: ON THE USE OF DYNAMIC BAYESIAN NETWORKS FOR SEMANTIC COMPOSITION

Marie-Jean Meurs, Fabrice Lefèvre and Renato de Mori

Université d'Avignon et des Pays de Vaucluse
Laboratoire Informatique d'Avignon (EA 931), F-84911 Avignon, France.
{marie-jean.meurs,fabrice.lefevre,renato.demori}@univ-avignon.fr

ABSTRACT

In the context of spoken language interpretation, this paper introduces a stochastic approach to infer and compose semantic structures. Semantic frame structures are directly derived from word and basic concept sequences representing the users' utterances. A rule-based process provides a reference frame annotation of the speech training data. Then dynamic Bayesian networks are used to hypothesize frames from test data. The semantic frames used in this work are specialized on the task domain from the Berkeley FrameNet set.

Experiments are reported on the French MEDIA dialog corpus. For all the data, the manual transcriptions and annotations at the word and concept levels are available. Tests are performed under 3 different conditions raising in difficulty wrt the errors in the word and concept sequence inputs. Three different stochastic models are compared and the results confirm the ability of the proposed probabilistic frameworks to carry out a reliable semantic frame annotation.

Index Terms— spoken dialog system, spoken language understanding, semantic frames, semantic composition, dynamic Bayesian networks.

1. INTRODUCTION

Stochastic methods are efficient alternatives to rule-based techniques for Spoken Language Understanding (SLU) [1, 2, 3]. In a spoken dialog system, the SLU module links up the automatic speech recognition (ASR) system and the dialog manager. After analysis of the user's utterance, it derives a representation of its semantic content from which the dialog manager can choose the next best action to perform, considering the current dialog context. Stochastic approaches can lower the need for human expertise requirement and reduce the development cost. Also they can produce lattices or *n*-best lists of hypotheses (with confidence scores) so as to carry on uncertainty up to the decision module.

Fully stochastic SLU models have already been proposed in former works [3, 4]. They are generally designed to improve the system robustness by progressively refining the hypothesized concept output. In this work, the objective is to introduce a richer semantic information in the system outputs in a relevant and adaptable way. To do so, an additional semantic composition step is considered so as to capture abstract semantics conveyed by the underlying basic concept representation. The meaning of the sentence constituents are composed to obtain an accurate and exhaustive representation of the meaning of the whole sentence. A frame formalism has been applied to well specify the structures used in this semantic composition step.

This work is supported by the 6th Framework Research Program of the European Union (EU), LUNA Project, IST contract no 33549, www.ist-luna.eu.

The semantic frames are defined according to the Berkeley FrameNet paradigm in which a frame describes a common or abstract situation involving pre-defined roles, named frame-elements (FE). The topic coverage of the FrameNet frames being generally too broad, specific frames and FE have been fixed, suited to the targeted dialog task [5]. A two-step rule-based process provides a semantic frame annotation of the speech data on top of the manual transcription and concept annotation. Though not perfect, this frame annotation has been shown to be quite reliable. However, as erroneous inputs must be considered, there is a need for a system able to produce n-best lists of hypotheses (or lattices) along with confidence scores which can be used by further validation steps.

In a recent paper [6], we reported some preliminary results on a system using the dynamic Bayesian networks (DBN) to handle this high level semantic composition step. Its performance showed that once a large enough corpus has been annotated in terms of semantic frames, it is possible to obtain the frame composition for a new utterance from a sequential frame decoding, even though long-span dependencies have been used to produce the training annotation. However in this previous work only frame heads were taken into account, not the FE.

The SLU system proposed in this paper is based on two decoding stages using DBN: a first stage derives basic concepts from user utterance transcription, then a second stage performs inferences on sequential semantic frame structures (including FE), considering all the available previous annotation levels (words and concepts). Three variants are presented and evaluated: a DBN model in which the frames and FE are jointly decoded as one variable, a model in which the frames and FE are split into two variables simultaneously decoded and eventually a 2-level model in which frames are decoded first then used as observed values for a FE decoding step.

The paper is organized as follows. The next Section presents the MEDIA corpus, reviews the background on semantic frames and describes the rule-based process used to provide the reference semantic frame annotation. Then Section 3 introduces the DBN-based models for semantic frame and FE composition and finally Section 4 reports on the experiments.

2. SEMANTIC FRAME ANNOTATION ON THE MEDIA CORPUS

The MEDIA corpus is a French dialog corpus simulating a tourist information phone server [7]. It has been recorded using a *Wizard of Oz* system. The corpus accounts 1257 dialogs spread over 70 hours of speech. Each of the 250 speakers recorded five different hotel reservation *scenarii*. The MEDIA corpus is manually transcribed and conceptually enriched with more than 80 basic concepts manually annotated. Semantic structures can be derived from semantic

```
<frame frname="LOCATION">
    <concept value="locate" />
    <lexical_units value="place, area" />
    <framelement fename="location_town">
      <concept value="town" />
      <lexical_units value="paris, marseille..." />
    </framelement>
    ...
</frame>
```

Table 1. Excerpt of the MEDIA frame LOCATION definition.

knowledge obtained with a semantic theory as semantic networks [8] or function/argument structures [9].

The semantic dictionary used to annotate the MEDIA corpus associates *concept-value* pairs to word segments. Indeed, this semantic dictionary contains *lexical* concepts as defined in Jackendoff's terminology [9]. These concepts, expressed by the words in the sentence, are the basic units out of which a semantic representation can be thought. To obtain a full representation of the semantic composition of an utterance, this work uses semantic frames. A semantic frame is a computational model representing semantic entities and their properties [10]. The choice of a frame annotation in this work is motivated by its ability to represent negotiation dialogs and also to adapt to complex actions of the dialog manager.

For a given frame or FE, the evoking words are its lexical units (LU). A LU is a pairing of a word with a meaning. The Berkeley FrameNet project [11] provides a large frame database for English, but no such database exists for French. Hence, we have manually defined a frame knowledge source to describe the semantic knowledge on the MEDIA domain. The MEDIA knowledge source contains 21 frames and 86 FE described by a set of manually defined patterns. These patterns are made of LU and conceptual units (CU). Some CU match the MEDIA basic concepts, others are defined according to the knowledge source frames. The example of the MEDIA frame LOCATION is given in Table 2, with one of its FE named location_town.

In order to obtain frame annotations of the speech data, a two-step rule-based annotation process has been carried out: firstly the patterns associated to frames are used to trigger new frames and their FE when they match with concept or word inputs, secondly a set of logical rules is applied to compose these frames. In the latter step, the frames and FE previously produced determine the truth values of the logical rules. According to these truth values, new frames and FE can be created and current frames and FE can be deleted, modified or connected. Some frames can be subframes of others, in this case they are connected through a FE taking a frame as value.

Figure 1 illustrates a Prolog logical rule:

```
do_link(LODH,H) :-
is_fe(lodging_hotel,LODH),
is_concept_of(hotel,LODH),
is_fr(HOTEL,H).
```

In this example, the rule creates a subframe link between the FE lodging_hotel and the frame HOTEL.

Around 70 rules are currently used. These rules do not depend neither on the words of the utterance nor on any sequentiality or order of appearance of the frames. This procedure allows to setup a reference frame annotation for the training corpus from which the stochastic models can be learned.

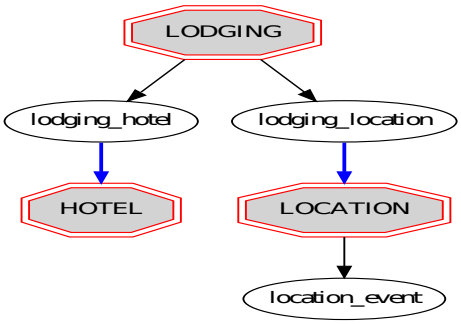

Fig. 1. Frames, FE and subframe relations associated to the word sequence *"staying in a hotel near the Festival d'Avignon"*

3. DBN-BASED FRAME COMPOSITION MODEL

The DBN framework offers a great flexibility for complex stochastic system representation. Lately, DBN have been used in many sequential data modeling tasks and generally state-of-the-art performance are observed [3]. Practical models used in the system are depicted in figures 2 and 3. For the sake of clarity, some additional vertices (*variables*) and edges (*conditional dependency*) are not represented and only two time slices (or two words) are depicted. In practice, a regular pattern is repeated until it fits the whole word sequence. Plain nodes are observed variables whereas empty nodes are hidden. Plain lines represent conditional dependencies between variables, dashed lines indicate switching parents (variables modifying the conditional relationship between others).

The left part of the figure 2 shows the generative DBN model in which the frames and FE are merged in one variable; in the right part, the frames and FE are associated to two variables but simultaneously decoded. The 2-level model in which frames are decoded first then used as observed values in the FE decoding step is shown in Figure 3.

The rationale for merging frames and FE into a single variable is the decoding complexity reduction: the set of possible frame/FE values is limited to the number of frame/FE combinations observed in the training data. However it leads to deterministic and unchangeable links between frames and FE. Factorization of frame and FE variables allows to deal with the ambiguities in the frame and FE links by adding probabilities to them and by testing every combination (even not encountered in the training data, by means of a backoff technique) during the decoding step. Obviously, it has a cost and the complexity of the model is so much increased that a sub-optimal beam search has to be used during decoding. Even though it is sub-optimal, the 2-level approach allows to reduce the complexity of the factored approach without loosing the model generalization improvement.

All variables being observed during training, the edge's conditional probability tables are directly derived from observation counts. To improve their estimates, factored language models (FLM) have been used along with generalized parallel backoff (GPB) [12]. FLM are an extension of standard language models in which the prediction is based upon a set of features (and not only on previous occurrences of the predicted variable). GPB allows to extend the standard backoff procedures to the case where heterogeneous feature types are considered and no obvious temporal order exists (contrary to classical language models, features in FLM can occur at the time of the prediction).

Several FLM implementations are used in the DBN models, corresponding to the arrows in the graph representations (Figures 2 and 3). In these FLM implementations, h represents the history length

($h = -1$ for 2-grams), FFE, F, FE, C and W respectively stand for frame/FE (one variable), frame, FE, concept and word variables. GPB uses the modified Kneser-Ney discounting technique in all conditions and works with order $\{i, j, ...\}$ in the calculation of the product $\prod P(u|i, j, ...)$. The conditional probability tables of the DBN models are provided by the FLM implementations as follows:

• Frame/FE compound variable:

$$P(FFE) \simeq \prod P(ffe|ffe_h);$$
$$P(C|FFE) \simeq \prod P(c|c_h, ffe);$$
$$P(W|C, FFE) \simeq \prod P(w|w_h, c, ffe).$$

• Frame and FE variables, simultaneous decoding:

$$P(F) \simeq \prod P(f|f_h);$$
$$P(FE|F) \simeq \prod P(fe|fe_h, f);$$
$$P(C|FE, F) \simeq \prod P(c|c_h, fe, f);$$
$$P(W|C, FE, F) \simeq \prod P(w|w_h, c, fe, f).$$

• Frame and FE variables, 2-level decoding:

– *First stage:*

$$P(F) \simeq \prod P(f|f_h);$$
$$P(C|F) \simeq \prod P(c|c_h, f);$$
$$P(W|C, F) \simeq \prod P(w|w_h, c, f).$$

– *Second stage:*

$$P(\hat{F}) \simeq \prod P(\hat{f}|\hat{f}_h);$$
$$P(FE|\hat{F}) \simeq \prod P(fe|fe_h, \hat{f});$$
$$P(C|\hat{F}, FE) \simeq \prod P(c|c_h, \hat{f}, fe);$$
$$P(W|C, \hat{F}, FE) \simeq \prod P(w|w_h, c, \hat{f}, fe).$$

The `word`, `concept` and `transition` sequences are observed variables for the frame and FE decoding: they have been decoded by the ASR and SLU modules. Due to data sparseness, the conditional probabilities used in the models are limited to 2-gram FLM.

Due to the frame hierarchical representation, some overlapping situations can occurred when determining the frame and FE associated to a concept. It arises mainly when several frames or FE can have been triggered by the same concept but also if inference and composition have created nested structures tied to the same concept. To address this difficulty a tree-projection algorithm has been developed and applied.

The projection is performed on the whole utterance tree-structured frame annotation and allows to derive sub-branches associated to a concept. Starting from a leaf of the tree, a compound frame/FE class is obtained by aggregating the father vertices (either frames or FE) as long as they are associated to the same concept (or none). The edges are defined by the frame→FE links and the FE→frame subframe relations [1]. For example, the word sequence "near the Festival d'Avignon", proposed in Figure 1, entails the creation of the projected branch: `location_event-LOCATION-lodging_location-LODGING`.

Thereafter, either the branches are kept unchanged and considered directly as compound classes (as in the frame/FE approach) or a separation is made between the frame and FE parts to produce two sets of classes (as in the 2-level approach). Compound frame and FE classes are considered in the decoding process then projected back afterward. The training corpus provides the set of frame and FE class sequences on which the DBN parameters are estimated.

[1] Cases exist where the frame annotation is not a tree. They have been found rare enough and in practice the proposed technique produces acceptable results for them.

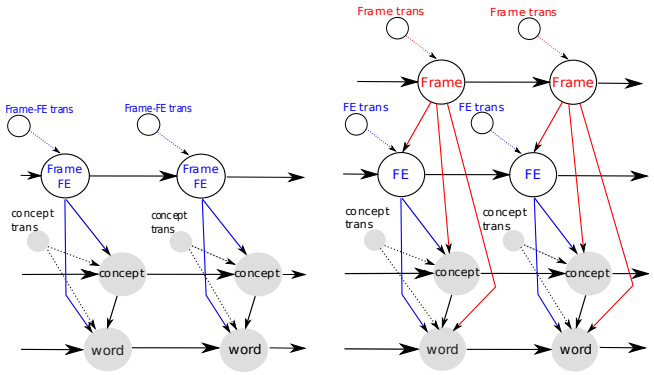

Fig. 2. Frames and FE as one or two unobserved variables

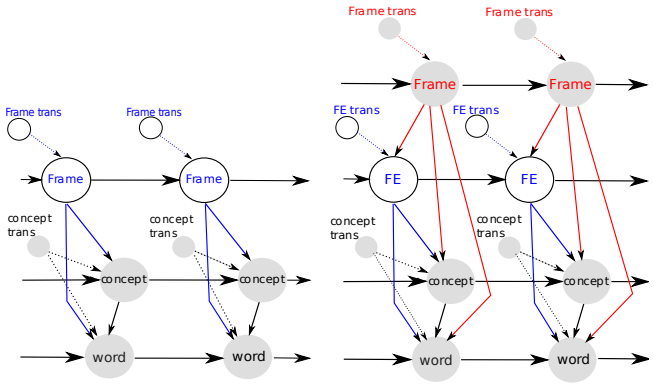

Fig. 3. 2-level decoding of frames and FE

4. EXPERIMENTS AND RESULTS

To evaluate the performance of the DBN-based frame composition systems, a test set is defined. Due to time and cost constraints, only 15 dialogs (containing 225 speaker turns) have been manually annotated with frames and FE by an expert. The two-step rule-based system (described in section 2) has been used to perform a frame annotation on the MEDIA data. The DBN model parameters have been trained on the training set using jointly the manual transcriptions, the manual concept annotations and the rule-based frame annotations.

Experiments are carried out on the test set under three different conditions according to the input type:

- REF (reference): the speaker turns are manually transcribed and annotated;
- SLU: the basic concepts are decoded from manual transcription of the speaker turns using a DBN-based SLU model conform to [13];
- ASR+SLU: word sequences are the 1-best hypotheses generated by an ASR system [14] and concepts are decoded using these hypotheses.

All the experiments reported in the paper have been performed using GMTK [15], a general purpose graphical model toolkit and SRILM [16], a language modeling toolkit.

In Table 2, the word and concept error rates are given for the 3 types of input. To serve as a baseline, the rule-based system is also evaluated on the test set. Table 2 is populated with the results on the test set for the rule-based and DBN-based frame composition

Systems		Inputs	REF		SLU		ASR + SLU	
		WER	0.0		0.0		14.8	
		CER	0.0		10.6		24.3	
			Frames	*FE*	Frames	*FE*	Frames	*FE*
rule-based		$\bar{p}$	0.93	*0.93*	0.92	*0.93*	0.89	*0.90*
		$\bar{r}$	0.98	*0.89*	0.97	*0.89*	0.87	*0.91*
		$\bar{F}$-m	**0.95**	***0.89***	**0.94**	***0.88***	**0.84**	***0.87***
DBN-based	frame/FE (compound variable)	$\bar{p}$	0.91	*0.91*	0.87	*0.91*	0.86	*0.90*
		$\bar{r}$	0.91	*0.84*	0.81	*0.82*	0.79	*0.86*
		$\bar{F}$-m	**0.89**	***0.85***	**0.81**	***0.81***	**0.78**	***0.84***
	frames and FE (2 variables)	$\bar{p}$	0.93	*0.93*	0.88	*0.92*	0.87	*0.90*
		$\bar{r}$	0.91	*0.84*	0.81	*0.83*	0.79	*0.86*
		$\bar{F}$-m	**0.89**	***0.85***	**0.80**	***0.83***	**0.78**	***0.83***
	2-level	$\bar{p}$	0.92	*0.91*	0.88	*0.91*	0.87	*0.89*
		$\bar{r}$	0.92	*0.81*	0.85	*0.79*	0.80	*0.81*
		$\bar{F}$-m	**0.90**	***0.83***	**0.83**	***0.80***	**0.79**	***0.79***

Table 2. Precision ($\bar{p}$), Recall ($\bar{r}$) and F-measure ($\bar{F}$-m) obtained on the MEDIA frame test set for the rule-based and DBN-based frame composition systems.

systems in terms of precision, recall and F-measure. Only the frame and FE identity is considered, neither the constituents it relies on nor the order matter.

The results in Table 2 show that the rule-based and DBN-based system performance are quite comparable in terms of precision, recall and F-measure (the observed odd is barely significant). Among the three DBN-based models, the results are slightly in favor of the 2-level approach, even though it is the most efficient model in term of decoding complexity. An average F-measure of 0.95 for the rule-based system on clean condition confirms that the semi-manual annotation process is quite reliable. After training on annotated data, DBN models are able to capture hierarchical structures and allow to obtain in one step what the human expert had to design in two. Also they can provide hypotheses with confidence scores and so be used in a n-best generation task or in a validation process.

5. CONCLUSION

In this paper, a stochastic process for generating and composing semantic frames using dynamic Bayesian networks has been investigated. The proposed approach offers a convenient way to automatically derive frame and frame-element annotations of speech utterances. Experimental results, obtained on the MEDIA dialog corpus, show that the performance of the DBN-based models are comparable to those of a hand-design rule-based approach.

The next step will be to use the n-best hypotheses from the ASR and SLU modules to derive n-best hypotheses of semantic frames and FE with confidence scores. Also the dialog context will be taking into account to improve the relevance of the frame and FE hypotheses wrt the dialog course.

6. REFERENCES

[1] E. Levin and R. Pieraccini, "Concept-based spontaneous speech understanding system," in *ESCA Eurospeech*, 1995.

[2] Y. He and S. Young, "Spoken language understanding using the hidden vector state model," *Speech Communication*, vol. 48(3-4), pp. 262–275, 2005.

[3] F. Lefèvre, "Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation," in *IEEE ICASSP*, 2007.

[4] H. Bonneau-Maynard and F. Lefèvre, "A 2+1-level stochastic understanding model," in *IEEE ASRU*, 2005.

[5] M.-J. Meurs, F. Duvert, F. Béchet, F. Lefèvre, and R. De Mori, "Semantic frame annotation on the french MEDIA corpus," in *LREC*, 2008.

[6] M.-J. Meurs, F. Lefèvre, and R. De Mori, "A bayesian approach to semantic composition for spoken language interpretation," in *ISCA Interspeech*, 2008.

[7] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, D. Mostefa, and the Media consortium, "Semantic annotation of the MEDIA corpus for spoken dialog," in *ISCA Eurospeech*, 2005.

[8] W.A. Woods, *What's in a Link: Foundations for Semantic Networks*, Bolt and 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 *SIGLEX Workshop: Why, What, and How?*, 1997.

[11] C.J. Fillmore, C.R. Johnson, and M.R.L. Petruck, "Background to framenet," *International Journal of Lexicography*, vol. 16.3, pp. 235–250, 2003.

[12] J. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," in *HLT-NAACL*, 2003.

[13] F. Lefèvre, "A dbn-based multi-level stochastic spoken language understanding system," in *IEEE Workshop on SLT*, 2006.

[14] L. Barrault, C. Servan, D. Matrouf, G. Linarès, and R. De Mori, "Frame-based acoustic feature integration for speech understanding," in *IEEE ICASSP*, 2008.

[15] J. Bilmes and G. Zweig, "The graphical models toolkit: An open source software system for speech and time-series processing," in *IEEE ICASSP*, 2002.

[16] A. Stolcke, "Srilm an extensible language modeling toolkit," in *IEEE ICASSP*, 2002.