Learning Bayesian Networks for Semantic Frame Composition in a Spoken Dialog System

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

A stochastic approach based on Dynamic Bayesian Networks (DBNs) is introduced for spoken language understanding. DBN-based models allow to infer and then to compose semantic frame-based tree structures from speech transcriptions. Experimental results on the French MEDIA dialog corpus show the appropriateness of the technique which both lead to good tree identification results and can provide the dialog system with n-best lists of scored hypotheses.

1 Introduction

Recent developments in Spoken Dialog Systems (SDSs) have renewed the interest for the extraction of rich and high-level semantics from users' utterances. Shifting every SDS component from hand-crafted to stochastic is foreseen as a good option to improve their overall performance by an increased robustness to speech variabilities. For instance stochastic methods are now efficient alternatives to rule-based techniques for Spoken Language Understanding (SLU) (He and Young, 2005; Lefèvre, 2007).

The SLU module links up the automatic speech recognition (ASR) module and the dialog manager. From the user's utterance analysis, it derives a representation of its semantic content upon which the dialog manager can decide the next best action to perform, taking into account the current dialog context. In this work, the overall objective is to increase the relevancy of the semantic information used by the system. Generally the internal meaning representation is based on flat concept sets obtained by either keyword spotting or conceptual decoding. In some cases a dialog act can be added on top of the concept set. Here we intend to consider an additional semantic composition step which will capture the abstract semantic structures conveyed by the basic concept representation. A frame formalism is applied to specify these nested structures. As such structures do not rely on sequential constraints, pure left-right branching semantic parser (such as (He and Young, 2005)) will not apply in this case.

To derive automatically such frame meaning representations we propose a system based on a two decoding step process using dynamic Bayesian networks (DBNs) (Bilmes and Zweig, 2002): first basic concepts are derived from the user's utterance transcriptions, then inferences are made on sequential semantic frame structures, considering all the available previous annotation levels (words and concepts). The inference process extracts all possible sub-trees (branches) according to lower level information (*generation*) and composes the hypothesized branches into a single utterance-span tree (*composition*). A hand-craft rule-based approach is used to derive the seed annotated training data. So both approaches are not competing and the stochastic approach is justified as only the DBN system is able to provide n-best lists of tree hypotheses with confidence scores to a stochastic dialog manager (such as the very promising POMDP-based approaches).

The paper is organized as follows. The next section presents the semantic frame annotation on the MEDIA corpus. Then Section 3 introduces the DBN-based models for semantic composition and finally Section 4 reports on the experiments.

Proceedings of NAACL HLT 2009: Short Papers, pages 61–64,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

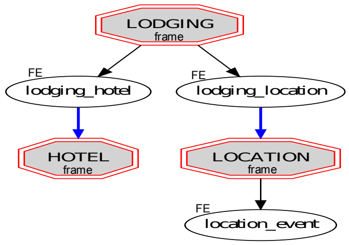

Figure 1: Frames, FEs and relations associated to the sequence *"staying in a hotel near the Festival de Cannes"*

2 Semantic Frames on the MEDIA corpus

MEDIA is a French corpus of negotiation dialogs among users and a tourist information phone server (Bonneau-Maynard et al., 2005). The corpus contains 1,257 dialogs recorded using a *Wizard of Oz* system. The semantic corpus is annotated with *concept-value* pairs corresponding to word segments with the addition of *specifier* tags representing some relations between concepts. The annotation utilizes 83 basic concepts and 19 specifiers.

Amongst the available semantic representations, the semantic frames (Lowe et al., 1997) are probably the most suited to the task, mostly because of their ability to represent negotiation dialogs. Semantic frames are computational models describing common or abstract situations involving roles, the frame elements (FEs). The FrameNet project (Fillmore et al., 2003) provides a large frame database for English. As no such resource exists for French, we elaborated a frame ontology to describe the semantic knowledge of the MEDIA domain. The MEDIA ontology is composed of 21 frames and 86 FEs. All are described by a set of manually defined patterns made of lexical units and conceptual units (frame and FE evoking words and concepts). Figure 1 gives the annotation of word sequence *"staying in a hotel near the Festival de Cannes"*. The training data are automatically annotated by a rule-based process. Pattern matching triggers the instantiation of frames and FEs which are composed using a set of logical rules. Composition may involve creation, modification or deletion of frame and FE instances. About 70 rules are currently used. This process is task-oriented and is progressively enriched with new rules to improve its accuracy. A reference frame annotation for the training corpus is established in this way and used for learning the parameters of the stochastic models introduced in the next section.

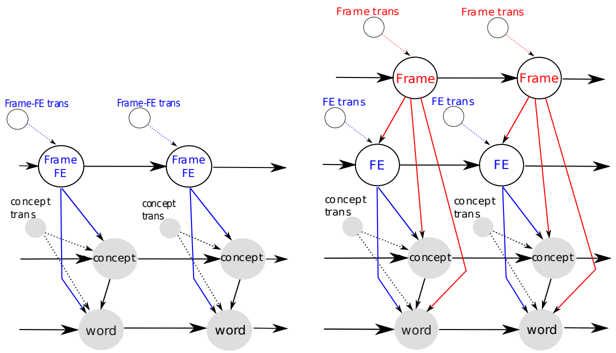

Figure 2: Frames, FEs as one or 2 unobserved variables

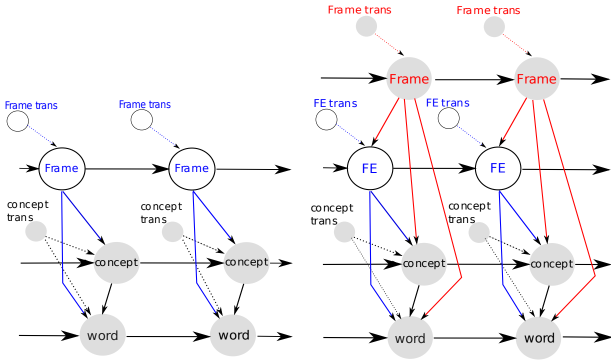

Figure 3: 2-level decoding of frames and FEs

3 DBN-based Frame Models

The generative DBN models used in the system are depicted on two time slices (two words) in figures 2 and 3. In practice, a regular pattern is repeated sufficiently to fit the entire word sequence. Shaded nodes are observed variables whereas empty nodes are hidden. Plain lines represent conditional dependencies between variables and dashed lines indicate switching parents (variables modifying the conditional relationship between others). An example of a switching parent is given by the `trans` nodes which influence the frame and FE nodes: when `trans` node is null the frame or FE stays the same from slice to slice, when `trans` is 1 a new frame or FE value is predicted based on the values of its parent nodes in the word sequence using frame (or FE) n-grams.

In the left DBN model of Figure 2 frames and FEs are merged in a single compound variable. They are factorized in the right model using two variables jointly decoded. Figure 3 shows the 2-level model where frames are first decoded then used as observed values in the FE decoding step. Merging frames and FEs into a variable reduces the decoding complexity but leads to deterministic links between frames and FEs. With their factorization, on the contrary, it is possible to deal with the ambiguities in the frame and FE links. During the decoding step, every combination is tested, even not encountered in the training data, by means of a back-off technique. Due to the increase in model complexity, a sub-optimal beam search is applied for decoding. In this way, the 2-level approach reduces the complexity of the factored approach while preserving model generalization.

Because all variables are observed at training time, the edge's conditional probability tables are directly derived from observation counts. To improve their estimates, factored language models (FLMs) are used along with generalized parallel backoff (Bilmes and Kirchhoff, 2003). Several FLM implementations of the joint distributions are used in the DBN models, corresponding to the arrows in Figures 2 and 3. In the FLMs given below, n is the history length ($n = 1$ for bigrams), the uppercase and lowercase letters FFE, F, FE, C and W respectively stand for frame/FE (one variable), frame, FE, concept and word variables:

- Frame/FE compound variable:

$$P(FFE) \simeq \prod_{k=0}^{n} P(ffe_k|ffe_{k-1});$$

$$P(C|FFE) \simeq \prod_{k=0}^{n} P(c_k|c_{k-1}, ffe_k);$$

$$P(W|C, FFE) \simeq \prod_{k=0}^{n} P(w_k|w_{k-1}, c_k, ffe_k).$$

- Frame and FE variables, joint decoding:

$$P(F) \simeq \prod_{k=0}^{n} P(f_k|f_{k-1});$$

$$P(FE|F) \simeq \prod_{k=0}^{n} P(fe_k|fe_{k-1}, f_k);$$

$$P(C|FE, F) \simeq \prod_{k=0}^{n} P(c_k|c_{k-1}, fe_k, f_k);$$

$$P(W|C, FE, F) \simeq \prod_{k=0}^{n} P(w_k|w_{k-1}, c_k, fe_k, f_k).$$

- Frame and FE variables, 2-level decoding:
  - *First stage:* same as frame/FE compound variables but only decoding frames
  - *Second stage:* same as joint decodind but frames are observed

$$P(\hat{F}) \simeq \prod_{k=0}^{n} P(\hat{f}_k|\hat{f}_{k-1});$$

$$P(FE|\hat{F}) \simeq \prod_{k=0}^{n} P(fe_k|fe_{k-1}, \hat{f}_k);$$

$$P(C|\hat{F}, FE) \simeq \prod_{k=0}^{n} P(c_k|c_{k-1}, \hat{f}_k, fe_k);$$

$$P(W|C, \hat{F}, FE) \simeq \prod_{k=0}^{n} P(w_k|w_{k-1}, c_k, \hat{f}_k, fe_k).$$

Variables with hat have observed values.

Due to the frame hierarchical representation, some overlapping situations can occurred when determining the frame and FE associated to a concept. To address this difficulty, a tree-projection algorithm is performed on the utterance tree-structured frame annotation and allows to derive sub-branches associated to a concept (possibly more than one). Starting from a leaf of the tree, a compound frame/FE class is obtained by aggregating the father vertices (either frames or FEs) as long as they are associated to the same concept (or none). The edges are defined both by the frame→FE attachments and the FE→frame sub-frame relations.

Thereafter, either the branches are considered directly as compound classes or the frame and FE interleaved components are separated to produce two class sets. These compound classes are considered in the decoding process then projected back afterwards to recover the two types of frame↔FE connections. However, some links are lost because decoding is sequential. A set of manually defined rules is used to retrieve the missing connections from the set of hypothesized branches. Theses rules are similar to those used in the semi-automatic annotation of the training data but differ mostly because the available information is different. For instance, the frames cannot anymore be associated to a particular word inside a concept but rather to the whole segment. The training corpus provides the set of frame and FE class sequences on which the DBN parameters are estimated.

4 Experiments and Results

The DBN-based composition systems were evaluated on a test set of 225 speakers' turns manually annotated in terms of frames and FEs. The rule-based system was used to perform a frame annotation of the MEDIA data. On the test set, an average F-measure of 0.95 for frame identification confirms the good reliability of the process. The DBN model parameters were trained on the training data using jointly the manual transcriptions, the manual concept annotations and the rule-based frame annotations.

Experiments were carried out on the test set under three conditions varying the input noise level:

- REF (reference): speaker turns manually transcribed and annotated;
- SLU: concepts decoded from manual transcriptions using a DBN-based SLU model comparable to (Lefèvre, 2007) (10.6% concept error rate);
- ASR+SLU: 1-best hypotheses of transcriptions

	Inputs		REF			SLU			ASR + SLU	
DBN models		Frames	*FE*	**Links**	Frames	*FE*	**Links**	Frames	*FE*	**Links**
frame/FEs	p̄/r̄	0.91/0.93	*0.91*/0.86	**0.93/0.98**	0.87/0.82	*0.91*/0.83	**0.93/0.98**	0.86/0.80	*0.90*/0.86	**0.92/0.98**
(compound)	**F̄-m**	**0.89**	***0.86***	**0.92**	**0.81**	***0.82***	**0.92**	**0.78**	***0.84***	**0.92**
frames and FEs	p̄/r̄	0.92/0.92	*0.92*/0.85	**0.94/0.98**	0.88/0.81	*0.92*/0.83	**0.93/0.97**	0.87/0.79	*0.90*/0.86	**0.94/0.97**
(2 variables)	**F̄-m**	**0.90**	***0.86***	**0.94**	**0.80**	***0.83***	**0.91**	**0.78**	***0.84***	**0.93**
frames then FEs	p̄/r̄	0.92/0.94	*0.91*/0.82	**0.92/0.98**	0.88/0.86	*0.91*/0.80	**0.92/0.97**	0.87/0.81	*0.89*/0.82	**0.93/0.98**
(2-level)	**F̄-m**	**0.91**	***0.83***	**0.93**	**0.83**	***0.80***	**0.90**	**0.79**	***0.80***	**0.92**

Table 1: Precision (p̄), Recall (r̄) and F-measure (F̄-m) on the MEDIA test set for the DBN-based frame composition systems.

generated by an ASR system and concepts decoded using them (14.8% word error rate, 24.3% concept error rate).

All the experiments reported in the paper were performed using GMTK (Bilmes and Zweig, 2002), a general purpose graphical model toolkit and SRILM (Stolcke, 2002), a language modeling toolkit.

Table 1 is populated with the results on the test set for the DBN-based frame composition systems in terms of precision, recall and F-measure. For the FE figures, only the reference FEs corresponding to correctly identified frames are considered. Only the frame and FE names are considered, neither their constituents nor their order matter. Finally, results are given for the sub-frame links between frames and FEs. Table 1 shows that the performances of the 3 DBN-based systems are quite comparable. Anyhow the 2-level system can be considered the best as besides its good F-measure results, it is also the most efficient model in terms of decoding complexity. The good results obtained for the sub-frame links confirm that the DBN models combined with a small rule set can be used to generate consistent hierarchical structures. Moreover, as they can provide hypotheses with confidence scores they can be used in a multiple input/output context (lattices and n-best lists) or in a validation process (evaluating and ranking hypotheses from other systems).

5 Conclusion

This work investigates a stochastic process for generating and composing semantic frames using dynamic Bayesian networks. The proposed approach offers a convenient way to automatically derive semantic annotations of speech utterances based on a complete frame and frame element hierarchical structure. Experimental results, obtained on the MEDIA dialog corpus, show that the performance of the DBN-based models are definitely good enough to be used in a dialog system in order to supply the dialog manager with a rich and thorough representation of the user's request semantics. Though this can also be obtained using a rule-based approach, the DBN models alone are able to derive n-best lists of semantic tree hypotheses with confidence scores. The incidence of such outputs on the dialog manager decision accuracy needs to be asserted.

Acknowledgment

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

References

J. Bilmes and K. Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *NAACL HLT*.

J. Bilmes and G. Zweig. 2002. The graphical models toolkit: An open source software system for speech and time-series processing. In *IEEE ICASSP*.

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

C.J. Fillmore, C.R. Johnson, and M.R.L. Petruck. 2003. Background to framenet. *International Journal of Lexicography*, 16.3:235–250.

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

F. Lefèvre. 2007. Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation. In *IEEE ICASSP*.

J.B. Lowe, C.F. Baker, and C.J. Fillmore. 1997. A frame-semantic approach to semantic annotation. In *SIGLEX Workshop: Why, What, and How?*

A. Stolcke. 2002. Srilm an extensible language modeling toolkit. In *IEEE ICASSP*.