interspeech 2008 incorporating SST 08

A Bayesian Approach to Semantic Composition for Spoken Language Interpretation

Marie-Jean Meurs, Fabrice Lefèvre, 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

10.21437/Interspeech.2008-352

Abstract

This paper introduces a stochastic interpretation process for composing semantic structures. This process, dedicated to spoken language interpretation, allows to derive semantic frame structures directly from word and basic concept sequences representing the users' utterances. First a two-step rule-based process has been used to provide a reference semantic frame annotation of the speech training data. Then, through a decoding stage, dynamic Bayesian networks are used to hypothesize frames with confidence scores from test data. The semantic frames used in this work have been derived from the Berkeley FrameNet paradigm.

Experiments are reported on the MEDIA corpus. MEDIA is a French dialog corpus recorded using a *Wizard of Oz* system simulating a telephone server for tourist information and hotel booking. For all the data the manual transcriptions and annotations at the word and concept levels are available. In order to evaluate the robustness of the proposed approach tests are performed under 3 different conditions raising in difficulty wrt the errors in the word and concept sequence inputs: (i) according to whether they are manually transcribed and annotated, (ii) manually transcribed and enriched with concepts provided by an automatic annotation, (iii) fully automatically transcribed and annotated. From the experiment results it appears that the proposed probabilistic framework is able to carry out semantic frame annotation with a good reliability, comparable to a semi-manual rule-based approach.

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

1. Introduction

Recently, stochastic techniques have been shown to be an efficient alternative to rule-based techniques for Spoken Language Understanding (SLU) [1, 2, 3, 4, 5]. They lower the need for human expertise and development cost and can provide lattices (or n-best) of hypotheses with confidence scores. Inside a spoken dialog system, the SLU module is the interface between the automatic speech recognition (ASR) system and the dialog manager. Its role is to analyze the user's query so as to derive a representation of its semantic content from which the dialog manager can decide its next best action to perform considering the current dialog context.

In former works [6], SLU systems in which the whole understanding process is stochastic have been proposed. In particular the baseline 2-level understanding system (such as in [2]) has been improved to a 2+1-level system through the integration of a stochastic value normalization phase which was formerly rule-based. In the stochastic approach, multi-stage SLU systems have already been proposed and investigated [4, 5]. However they are generally designed with the objective to improve the system robustness by progressively refining the hypothesized concept output. Our objective here 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 must be considered so as to capture abstract semantics convey by the underlying basic concept representation.

No general agreement exists on what the semantic structures should be in a spoken dialog system. We chose to use a frame formalism and to bound our frame definitions to the Berkeley FrameNet paradigm. Semantic frame structures have been retained for their ability to represent negotiation dialogs and also to adapt to complex actions of the dialog manager. A frame describes a common or abstract situation involving predefined roles. The topic coverage of the FrameNet frames being generally too broad, more specific frames have been defined, suited to the targeted dialog task [7]. A (semi-manual) two-step rule-based process has been developed and has allowed to provide a semantic frame annotation of the speech data on top of the manual transcriptions and concept annotation. This frame annotation while not perfect is quite reliable. However as erroneous inputs are to 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 this outlook, the proposition studied in this paper is to develop a SLU system based on two decoding stages using dynamic Bayesian networks (DBN). The first standard decoding stage derives basic concepts from user utterance transcription (such as in [5]). Then in a second stage, a DBN-based model performs inferences on sequential semantic structures, taking into account all the previous annotation levels available (words and concepts). Our assumption is 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 in the first place.

The paper is organized as follows. The next section presents the MEDIA corpus. Section 3 reviews background on semantic frames and describes the rule-based process used to provide the reference semantic frame annotation on the MEDIA corpus. Then Section 4 introduces the DBN-based model for semantic frame composition and finally Section 5 reports on the experiments.

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.

W^c	*concept c*	*mode*	*specifier*	*value*
I'd like to book	`command`	+		reservation
a room	`room-amount`	+	reservation	1
for two nights	`night-amount`	+	reservation	2
in Marseille	`location-town`	+	hotel	Marseille

Table 1: Example of the MEDIA semantic annotation.

2. MEDIA Corpus

The MEDIA corpus is a French dialog corpus simulating a telephone server for tourist information and hotel booking [8]. It has been recorded using a *Wizard of Oz* system. Eight *scenarii* categories were defined with various complexity levels. The corpus accounts 1257 dialogs from 250 speakers and contains about 70 hours of speech. Each speaker recorded five different hotel reservation *scenarii*. The MEDIA corpus is manually transcribed and conceptually enriched with more than 80 basic concepts manually annotated.

The semantic dictionary used to annotate the MEDIA corpus associates a *concept-value* pair to a word segment then a *specifier* showing the relations between concepts and also a *mode* (positive, negative, interrogative or optional) attached to the concept. By defining a set of 19 *specifiers* which are combined with the basic concepts, the MEDIA annotation scheme preserves the relationships between concepts. It makes it possible to build a hierarchical representation of an utterance interpretation.

Table (1) gives an example of the MEDIA annotation for the message (translated from French) *" I'd like to book a room for two nights in Marseille"*. In this example, the `reservation` specifier is given to the `room-amount` and `night-amounts` concepts as a hierarchical structure representing a reservation is triggered by the concept `command` and filled with the elements found in `room-amount` and `night-amount`. The specifier *hotel* associated to the *location-town* concept connects the town named in the segment *" in Marseille"* with the previous part of the utterance. The combination of the specifiers and the attribute names allows recomposing a hierarchical representation of a query from its flat annotation. This annotation provides labels comparable to semantic constituents hypothesized by a semantic shallow parser. However, if one intend to obtain a full representation of the semantic composition of an utterance based on the basic building blocks, specifiers are too simple and more complex structures have to be sought.

3. Semantic Frame Annotation

Semantic structures can be derived from semantic knowledge obtained with a semantic theory. Examples are semantic networks to represent entities and their relations [9] or function/argument structures [10]. A semantic frame is a computational model representing semantic entities and their properties [11].

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. A frame describes a common or abstract situation involving roles called frame elements (FE). For a given frame, the frame-evoking words are its lexical units (LU). A LU is a pairing of a word with a meaning. The Berkeley FrameNet project [12] provides a frame database for English. It currently contains more than 10,000 LU, over 6,100 of which are fully annotated, in nearly 825 hierarchically-related semantic frames, exemplified in more than 135,000 annotated sentences.

The FrameNet dictionary is for English but no such database exists for French. Hence, we have manually defined a frame knowledge source (KS) to describe the semantic composition knowledge on the MEDIA domain. The MEDIA KS contains 21 frames and 86 FE. Frames and FE are described by a set of manually defined patterns. These patterns are made of LU, conceptual units (CU) and words (features extracted from the compounds of them can also be considered). Some of the CU match the MEDIA basic concepts, some others are defined according to the KS frames. The example of the MEDIA frame `LOCATION` with one of its FE named `location_town` is given in Table 2.

In order to obtain frame annotations on 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 produced in the first step 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 (for instance some frames can be subframes of others, in this case they are connected through an FE taking a frame as value).

Prolog [13] has been retained to perform the logical inferences as it is certainly the most widely used language for logic programming. Based on the mathematical notions of relations and logical inference, a Prolog program consists of a database of facts and logical rules describing the relationships between potential facts. An example of a Prolog rule is given in Table 3. In this example, a "subframe" link/relation is created between the `Reservation_Theme` FE and the `LODGING` frame (every FE name includes a reference to its containing frame, so no need to mention it in the rule: `Reservation_Theme` is an FE of the `RESERVATION` frame).

Approximately 70 rules are currently used in the process. The rules do not depend neither on the words of the utterance nor on any sequentiality or order of appearance of the frames. They mainly consist in creating links between frames and FE, instantiating frames and FE not discovered by pattern matching and also avoiding redundancies. The logical inference is apply iteratively (up to 5 times max, so as to keep computation time reasonable), each of its outputs providing the inputs of the next resolution search.

This procedure allows to setup a reference frame annotation for the training corpus from which the stochastic models can be learned.

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

Table 2: Excerpt of the MEDIA frame LOCATION definition.

```
do_link(RESL,L) :-  is_fe(reservation_theme,RESL),
                    is_concept_of(lodging,RESL),
                    is_fr(lodging,L).
```

Table 3: A Prolog rule linking the LODGING frame and the Reservation_Theme FE.

4. DBN-based Frame Composition Model

The dynamic Bayesian network framework offer a great flexibility for complex stochastic system representation. Lately, DBN have been used in many sequential data modeling tasks (ASR, POS and dialog-act tagging, DNA sequence analysis...). And generally state-of-the-art performance are observed.

Figure 1 shows the generative DBN model in the case of a semantic composition SLU system. For the sake of simplicity, some additional vertices (*variables*) and edges (*conditional dependency*) of the actual DBN used in the system are not represented. In the figure, 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). An example of a switching parent is given by the transition node which influences the frame node: when transition is null, frame is a mere copy of the previous frame but if it is set to 1 the new frame value is determined accordingly to the probability $P(f|f_{-1})$ of the frame f given the previous frame f_{-1}.

All variables are observed during training, so no EM training iterations are necessary. The edge's conditional probability tables can be directly derived from observation counts. To improve their estimates, factored language models (FLM) have been used along with generalized parallel backoff (GPB) [14]. FLM are an extension of standard LM where 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 LM, features in FLM can occur at the time of the prediction).

Several FLM implementations are used in the DBN frame model, corresponding to the arrows in the DBN graph representation (see figure 1):

- $P(F) \simeq \prod P(f|f_h)$: frames sequences;
- $P(C|F) \simeq \prod P(c|c_h, f)$, GPB works with order $\{c_h, f\}$: concept sequences conditioned on frames;

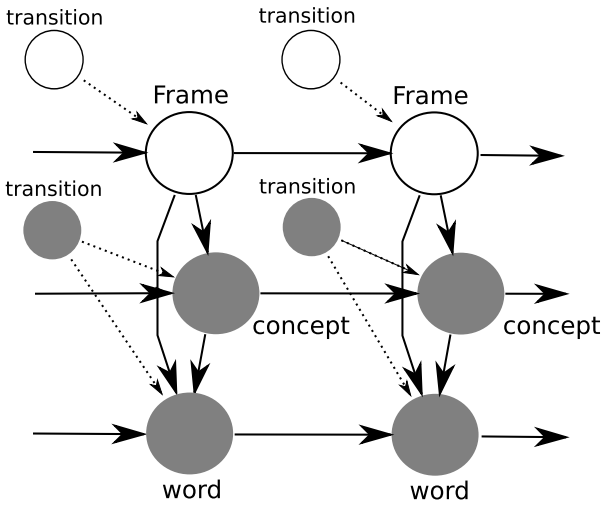

Figure 1: DBN-based semantic frame model. The model uses concept and word sequences as observation inputs for semantic frame decoding.

- $P(W|C, F) \simeq \prod P(w|w_h, c, f)$, GPB works with order $\{w_h, c, f\}$: word sequences conditioned on concepts and frames

where h represents an history which could vary according to the length of the model used ($\{-1\}$ for 2-grams, $\{-1, -2\}$ for 3-grams etc). GPB uses the modified Kneser-Ney discounting technique in all conditions. 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.

The DBN frame model used in the system is depicted in Figure 1. To start with, only frames are decoded (i.e. FE are not considered). To take into account the overlapping situations (where several frames can be associated to the same words or concepts) compound frame classes are considered in the decoding process and separated afterwards. The word, concept and transition sequences are observed variables for the frame decoding: they have been decoded by the ASR and SLU modules. Due to data sparseness, the conditional probabilities used in the model are limited to 2-grams FLM.

5. Experiments and Results

To evaluate the performance of the DBN-based frame composition system, a test set is defined. For time and cost purposes, only 15 dialogs (containing 225 speaker turns) have been manually annotated with frames by an expert. The two-step rule-based system (described in 3) has been used to perform a frame annotation on the MEDIA data (test set excluded). The FLM used in the DBN model have been trained on this 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:

- 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

Systems	Inputs	REF	SLU	ASR + SLU
	WER	0.0	0.0	14.8
	CER	0.0	10.6	24.3
	P	0.94	0.92	0.88
rule-based	R	0.92	0.87	0.80
	F-m	0.93	0.89	0.84
	P	0.91	0.91	0.82
DBN-based	R	0.89	0.79	0.76
	F-m	0.90	0.85	0.79

Table 4: Precision (P), Recall (R) and F-measure (F-m) obtained on the MEDIA frame test set for the rule-based and DBN-based frame composition systems. Word and concept error rates (%) on the test set considering 3 conditions for concept decoding: reference (REF), from manual transcription (SLU) and from ASR 1-best hypothesis (ASR+SLU).

model conform to [17];

- ASR+SLU: word sequences are the 1-best hypotheses generated by an ASR system [18] and concepts are decoded using these hypotheses.

In Table 4, 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 4 is populated with the results on the test set for the rule-based and DBN-based frame composition systems in terms of precision, recall and F-measure. Precision is defined as the number of correct semantic frames hypothesized by a system divided by the total number of hypothesized frames, recall is defined as the number of correct frames hypothesized by a system divided by the total number of reference frames. In both cases, only the frame identity is considered. Neither the constituents it relies on nor its order matter. The F-measure is the standard weighted harmonic mean of precision and recall ($\beta = 1$).

The results in Table 4 show that the DBN-based system performs comparably to the rule-based system in terms of precision, recall and F-measure. The slight difference between both systems (around 0.05 in F-measure) remains constant in the 3 contrastive conditions. The test set being small for the moment, the confidence interval radius is about 0.03 so the observed differences are not statistically very significant. A F-measure of 0.93 for the rule-based system on clean condition confirms that the semi-manual annotation process is quite reliable. Moreover, it is worth noting that the DBN-base system allows to obtain in one step what the human expert had to design in two: first hypothesizing a frame set from pattern-matching on the word and concept sequences then composing them to take into account their long-span relations at the utterance level.

6. Conclusion

In this paper, a stochastic interpretation process for composing semantic frames using dynamic Bayesian networks has been introduced. This process, dedicated to spoken language interpretation, allows to derive semantic frames from word sequences and basic concepts. Experimental results, obtained on the MEDIA dialog corpus, show that the performance of the DBN-based model are comparable to those of a hand-design rule-based approach.

The proposed approach offers a convenient way to automatically derive frame annotations of speech utterances. The next step will be to enrich the DBN-model by taking into account the FE. Due to the great flexibility in terms of probability representation of the DBN, it will merely consists in adding a new variable with the appropriate conditional probabilities in the graph. Also the n-best hypotheses from the ASR and SLU modules will be used to derive n-best hypotheses of semantic frames with their confidence scores.

7. References

[1] E. Levin and R. Pieraccini, "Concept-based Spontaneous Speech Understanding System", ESCA Eurospeech, 1995.

[2] F. Pla et al, "Language Understanding using Two-level Stochastic Models with POS and Semantic Units", LNCS series, vol. 2166, p. 403-409, 2001.

[3] Y. He and S. Young, "Spoken Language Understanding using the Hidden Vector State Model", Speech Communication, Vol. 48(3-4), p. 262-275, 2005.

[4] C. Raymond et al, "On the use of finite state transducers for semantic interpretation", Speech Communication, Vol 48:3-4, 288-304, 2006.

[5] F. Lefèvre, "Dynamic Bayesian Networks and Discriminative Classifiers for Multi-Stage Semantic Interpretation", IEEE ICASSP, 2007.

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

[7] M.-J. Meurs et al, "Semantic Frame Annotation on the French MEDIA corpus", LREC, 2008.

[8] H. Bonneau-Maynard et al, "Semantic annotation of the MEDIA corpus for spoken dialog", ISCA Eurospeech, 2005.

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

[10] R. Jackendoff, "Semantic Structure", The MIT Press, Cambridge Mass., 1990.

[11] J.B. Lowe and C.F. Baker and C.J. Fillmore, "A frame-semantic approach to semantic annotation", SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, 1997.

[12] C.J. Fillmore et al, "Background to FrameNet", International Journal of Lexicography, 2003.

[13] A. Colmerauer and P. Roussel, "The birth of Prolog", HOPL-II: The second ACM SIGPLAN conference on History of programming languages, p.37–52, 1993.

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

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

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

[17] F. Lefèvre, "A DBN-based multi-level stochastic spoken language understanding system", IEEE Workshop on SLT, 2006.

[18] L. Barrault et al, "Frame-Based Acoustic Feature Integration for Speech Understanding", IEEE ICASSP, 2008.