DYNAMIC BAYESIAN NETWORKS AND DISCRIMINATIVE CLASSIFIERS FOR MULTI-STAGE SEMANTIC INTERPRETATION

Fabrice Lefèvre

LIA - University of Avignon
fabrice.lefevre@univ-avignon.fr

ABSTRACT

In this paper, a multi-stage spoken language understanding system is presented. This stochastic module is for the first time based on a combination of dynamic Bayesian networks and conditional random field classifiers. The former generative models allow to derive basic concept sequences from the word sequences which are in turn augmented with modalities and hierarchical information by the latter discriminative models. To provide efficiently smoothed conditional probability estimates, factored language models with a generalized parallel backoff procedure are used as the network edge implementation. This framework allows a great flexibility in terms of probability representation facilitating the development of the stochastic levels (semantic and lexical) of the system.

Experiments are carried out on the French MEDIA task (tourist information and hotel booking). The MEDIA 10k-utterance training corpus is conceptually rich (more than 80 basic concepts) and is provided with a manually segmented annotation. On this complex task, the proposed multi-stage system is shown to offer better performance than the MEDIA'05 evaluation campaign best system [1].

Index Terms— spoken language understanding, dynamic Bayesian networks, conditional random field, multi-stage system

1. INTRODUCTION

In the research community, recent years have witnessed the emergence of stochastic techniques for spoken language understanding (SLU) as an efficient alternative to rule-based techniques by reducing the need for human expertise and development cost [2, 3, 4, 5]. In a spoken dialog system, the SLU module acts as an interface between the automatic speech recognition system and the dialog manager. The user query is analyzed and a representation of its semantic content is derived so that the dialog manager can take a context-sensitive decision about the dialog follow-up.

In former works [6], SLU systems have been proposed in which the understanding process is entirely stochastic. In particular the baseline 2-level understanding system (such as in [3]) has been improved to a 2+1-level system with 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 [5]. However they are generally designed with the objective to improve the system robustness by progressively refining the hypothesized output. In this work, the objective is to introduce richer information in the system outputs in a progressive and reliable way.

In this outlook, we propose to develop a multi-stage SLU system based on a first decoding stage by means of dynamic Bayesian networks (DBN) followed by several classification stages based on conditional random field (CRF) models. The first decoding stage derives a basic concept sequence which is subsequently augmented with additional information such as modalities, hierarchical dependencies... When all the possible semantic units are considered (combining an attribute, a modality, a value and some query-level dependencies), data sparseness becomes a crucial issue for semantic interpretation. On the one hand, discriminative classifiers are required to obtain good performance for very small-sized class detection. On the other hand, from our experience, these classifiers are generally not the most appropriate for sequential decoding processes (combined segmentation and identification). The foreseen advantage of the multi-stage approach is to improve the class modelization during the maximum-likelihood decoding (by ensuring a sufficient amount of examples per merged class). Then maximum *a posteriori* classification models are used to segregate the units inside the merged classes.

The MEDIA corpus [7] is used to evaluate the approach on a new and challenging task. The MEDIA task is about hotel room reservation along with tourist information in France, using information obtained from a web-based database. The corpus has been issued for the French Technolangue EVALDA-MEDIA understanding evaluation campaign [1] (6 participants) which has allowed in return to settle reference performance on the test data. The retained evaluation paradigm relies on a common generic semantic representation based on an attribute-value structure in which conceptual relationships among constituents are implicitly represented by the use of specifiers in the name of the attributes. To allow the comparison between SLU systems, only context-independent evaluation is considered: the SLU module works on isolated utterances without taking into account the dialog context.

The paper is organized as follows. The next section briefly reviews background on 2+1-level stochastic SLU model. Section 3 gives a description of the MEDIA corpus. Then the DBN-based SLU models are introduced in Section 4 followed by a presentation of the CRF classifiers in Section 5. Finally, the last section reports on the experimental results.

2. 2+1-LEVEL SLU MODEL

A main assumption for stochastic speech understanding (as formulated in [2]) is that there is a sequential correspondence between the word sequence and the underlying concept sequence. Be $W = w_1 w_2 \dots w_N$ the sequence of words in the sentence, the understanding process build the sequence of concepts $C = c_1 c_2 \dots c_N$ which maximizes the *a posteriori* probability, rewritten according to the Bayes formula:

$$\hat{C} = \arg\max_C P(C|W) = \arg\max_C P(W|C)P(C) \qquad (1)$$

By concept it is meant a combination of an attribute name and its modality (*affirmative*, *negative*...). The term $P(W|C)$ is estimated

words	mode	attribute name	normalized value
donnez-moi	+	`null`	
le	?	`refLink-coRef`	`singular`
tarif	?	`object`	`payment-amount-room`
puisque	+	`connectProp`	`imply`
je voudrais	+	`null`	
une chambre	+	`number-room`	`1`
qui coûte	+	`object`	`payment-amount-room`
pas plus de	+	`comparative-payment`	`less than`
cinquante	+	`payment-amount-integer-room`	`50`
euros	+	`payment-unit`	`euro`

Fig. 1. Semantic concept (att./value) representation for the query *"give me the rate for I'd like a room charged not more than fifty euros"*.

by means of n-gram probabilities of words given the concept associated to the predicted word and $P(C)$ is estimated in terms of m-gram probabilities of concepts .

Based on this formulation, several approaches can be considered depending on the orders of the models used to produce the estimates of $P(W|C)$ and $P(C)$. Generally, concept 2-grams $P(c_i|c_{i-1})$ are used to model the concept sequences. Associated to word 1-grams $P(w_i|c_i)$, this constitutes the basic model. Depending on the availability of training data with a segmental annotation, word 2-grams (or higher) conditioned on concept $P(w_i|w_{i-1}, c_i)$ can be used, leading to a 2-level model.

Traditionally, after a concept sequence has been decoded, the segmented word substrings are converted to a normalized form, as defined in the task semantic dictionary. In Figure 1, the normalization module translates the sequence *pas plus de* (*no more than*) assigned to the attribute `comparative-payment-room` to the normalized form `less than`. Many lexical sequences can correspond to the same normalized value. This normalization step, commonly based on manual rules, can be introduced in the global stochastic model through an additional level.

In this context, a new formulation of the concept decoding is applicable where the concept sequence is combined with the value sequence:

$$\hat{C}, \hat{V} = \arg\max_{C,V} P(C, V|W) \quad (2)$$

$$= \arg\max_{C,V} P(W|C, V)P(V|C)P(C) \quad (3)$$

As a consequence, the word sequence probabilities become conditioned both on the concepts and normalized values. The complexity of the conceptual model is then greatly increased (and the need for a sub-optimal decoding setup leads to poor performance [8]). To tackle this problem, the normalization level is not totally embodied in the conceptual model but instead delayed after the concept decoding. So, in the 2+1-level model, V is first marginalized then V is decoded with a constant C:

$$\hat{C} = \arg\max_{C} \sum_{V} P(W|C, V)P(V|C)P(C) \quad (4)$$

$$\hat{V} = \arg\max_{V} P(W|\hat{C}, V)P(V|\hat{C})P(\hat{C}) \quad (5)$$

Under the assumption that the normalized values have a slight or no influence on the segmentation process, Eq. 4 allows for a better generalization of the conceptual model.

3. THE MEDIA CORPUS

The MEDIA dialog corpus [1] has been recorded using a WOZ system simulating a vocal tourist information phone server [7]. 1257 dialogs were recorded, from 250 different speakers. The corpus is on the order of 70 hours of transcribed dialogs. The corpus consists in a training portion of 10965 requests, and a test portion of 3003 requests. An example of a semantic representation of a client utterance is given in Figure 1. A semantic segment is represented by a triplet which contains: the mode (+, -, ? and ~ meaning `optional` as in *with a shower if possible*), the attribute name and the normalized value. In order to disambiguate sentences such as *not in Paris in Nancy*, modes are assigned in a per segment basis. The order of the triplets in the semantic representation follows their order in the utterance.

The attribute values are either numeric units, proper names or semantic classes merging lexical units which are synonyms for the task. The set of normalized values associated to each attribute are defined in the semantic dictionary with 3 different possible configurations: a value list (eg `comparative` with possible values `around`, `less-than`, `maximum`, `minimum` and `more-than`), regular expressions (as for dates) or open values (i.e. no restrictions, as for client names). In the MEDIA annotation scheme the relationships between segments are preserved by defining a set of specifiers which are combined with basic attributes. In Figure 1, the attribute `comparative-payment` is derived from the combination of the `comparative` attribute and the `payment` specifier and the attribute `payment-amount-integer-room` is derived from the combination of the `payment-amount-integer` attribute with the specifier `room`. This combination of attribute names and specifiers is comparable to a semantic shallow parsing and makes it possible to derive a hierarchical representation of a query from its flat annotation [7].

A total of 676 proper names appear in the corpus, mainly city names (201) and hotel names (548) which are highly ambiguous for the task. The semantic dictionary includes 83 basic attributes and 19 specifiers. The combination of the basic attributes and the specifiers results in 1121 potential attributes. A total of 144 distinct attributes appear in the training corpus (and around 2.2k different normalized values.

4. DBN-BASED UNDERSTANDING MODELS

Over the last years, dynamic Bayesian networks have been investigated for many sequential data modeling tasks (automatic speech recognition, POS and dialog-act tagging [9], DNA sequence analy-

[1] soon available at http://www.elda.org

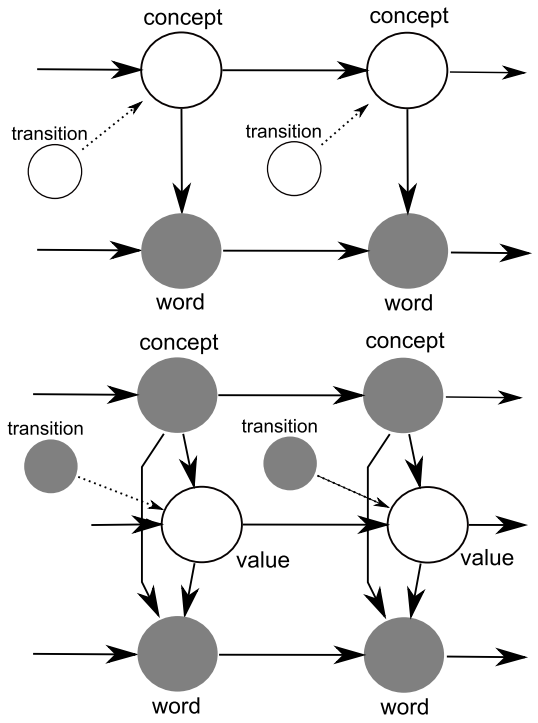

Fig. 2. DBN-based 2+1-level SLU system. Concept model (upper graph) is used for concept decoding. Value model (lower graph) uses concept sequences as observations for value identification.

sis...). DBN have shown a great flexibility for complex stochastic system representation and good performance are generally observed when compared to other standard techniques.

Figure 2 shows two generative DBN models in the case of a 2+1-level SLU system. For the sake of simplicity, some additional vertices (*variables*) and edges (*conditional dependency*) of the actual DBNs used in the system are not represented. In the figure, only two time slices (or two words) are depicted. In practice, the regular pattern (`chunk`) is repeated so as to fit the entire word sequence under consideration. Filled nodes are observed variables whereas blank ones are hidden. Plain lines represent conditional dependencies between variables, dashed lines indicate switching parents (variables influencing the relationship between others). An example of a switching parent is given by the `transition` node which influences the `concept` node: when `transition` is null, `concept` is a mere copy of the previous concept but when it is set to 1 the new `concept` value is determined accordingly to $P(c|c_{-1})$.

In our context, all variables are observed during training and so no EM iterations are necessary to learn parameters. The edge's conditional probability tables can be directly derived from observation counts. However, in order to improve their estimates, factored language models (FLM) have been used along with generalized parallel backoff (GPB) [10]. 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). To complement this new framework, 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 SLU models, each one of them corresponding to an arrow in the DBN graph representations (see figure 2):

- $P(C) \simeq \prod P(c|c_h)$: concept sequences;
- $P(V|C) \simeq \prod P(v|c)$: values conditioned on concepts;
- $P(W|C) \simeq \prod P(w|w_h, c)$, GPB works with order $\{w_h, c\}$: word sequences conditioned on concepts;
- $P(W|C, V) \simeq \prod P(w|w_h, v, c)$, GPB works with order $\{w_h, c, v\}$: word sequences conditioned on concepts and values

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 [11], a general purpose graphical model toolkit and SRILM [12], a language modeling toolkit.

The DBN models used in the system are depicted in Figure 2. The `concept` and `value` decoding steps are decoupled and correspond to the upper and lower graphs in Figure 2. The conceptual decoding outputs (`concept` and `transition` sequences) become observed variables for the value decoding. Conditional probabilities are either 2 or 3-grams FLM.

5. DISCRIMINATIVE MODELS

The DBN-based models proposed above present two major flaws: 1) they have no internal mechanism to deal with hierarchical and long-span dependencies such as the ones implies by the concept specifiers, 2) they are not discriminant which becomes problematic when data sparsity is as high as you have many singleton classes.

Even though techniques applicable to sequential stochastic models have been recently proposed [4], a full tree-based semantic parsing of the sentences seems out of the scope of models based on n-grams. So to derive hierarchical and query-level dependencies, it has been chosen to add the specifiers after the decoding stage by means of a discriminative classifier applied on the hypothesized interpretations. The total number of classes considered during decoding is greatly reduced, leading to a better generalization of the models. The same method has been applied to disambiguate the 4 modalities: only the two most represented (*affirmative* and *negative*) are considered during decoding and the others are mapped to *affirmative*.

Among the competitive approaches for discriminative classification (perceptron, CRF, SVM,...), CRF [14] have been retained in our experiments considering their good performance proved in several works (e.g. [13]). A first CRF model is trained to detect when appropriate specifiers should be added to a concept. The features used by the CRF model are the concept neighbors (both in the word and concept sequences) and 3 triggers associated to the main objects of the task (reservation, hotel, room). These triggers detect the presence of keywords in the utterance (such as *reservation, room*...) and thus provide a simple way to take into account long-span dependencies. A second CRF model projects back the *affirmative* modality to either *affirmative, interrogative* or *optional*. Its features are also the neighbors in the word and concept sequences and the hypothesized mode (which allows to take into account the negative mode during the classification process). The toolkit `CRF++` [2] has been used in the experiments reported below.

[2] available at http://chasen.org/ taku/software/CRF++/

decoding models		4 modes	specifiers	*relax*		*full*	
				2 modes	4 modes	2 modes	4 modes
HMM 2+1-level	2g	in HMM	manual rules	21.6	27.0	23.8	29.0
DBN 2+1-l	2g	in DBN	CRF	21.3	27.0	23.9	30.5
DBN 2+1-l	3g	in DBN	CRF	20.8	27.9	23.3	29.9
DBN 2+1-l	2g	CRF	CRF	20.0	25.3	22.1	27.4
DBN 2+1-l	3g	CRF	CRF	19.4	24.5	21.6	26.6

Table 1. Understanding error rates (%) on the MEDIA test set for 3 different systems with different evaluation modes and n-gram order (2 and 3-grams).

6. EXPERIMENTS AND RESULTS

The scoring tool developed for the MEDIA project allows to align two semantic representations and to compare them in terms of deletion, insertion, and substitution. The scoring can be done on the whole triplet including [mode, attribute name and attribute value] (*full* eval). It is also performed by using simplifications which consist in applying: 1) a projection on modes ' ~ ', and '?' to '+' (*2 mode* eval, '+' and '-'), and 2) a relaxation function on the attribute names removing the hierarchical specifiers (*relax* eval).

The baseline HMM-based SLU system relies on the methodology described in [6]. The 2+1-level conceptual model is trained on the 10,965-utterance training corpus. Several lexical classes (hotel names, city names, contry...) are used to generalize the estimate of the 2-grams. In order to deal with ambiguities (a same sequence of words can belong to several classes depending on the underlying concept), the classes are applied selectively to the concepts. No such generalization technique has been applied to the DBN-based systems. However in all the experiments, a specific treatment is done for the normalization of numbers and dates. Also *filler* words (such as *euh, ah*...) are removed from the transcriptions.

The understanding error rates for the reference HMM system are shown in Table 1: 29.0% in *full* eval with 4 modes, down to 21.6% in *relax* eval with 2 modes [1]. The DBN 2+1-level system offers performance comparable to the HMM system in *relax* eval (no specifiers) and a slight decrease in *full* eval. We observe that the modality classification is more complex ($\simeq$ 5% loss between 2 and 4 modes) than specifiers attribution ($\simeq$ 2% loss). The multi-stage approach allows to fill the gap by improving the modality classification, leading to a 11% relative improvement in *full* mode (from 29.9% to 26.6%).

Shifting from 2-grams to 3-grams has been considered in the DBN simultaneously at the concept level ($P(C)$) and the word level ($P(W|C)$ and $P(W|C,V)$). Globally it allows a 0.8% error rate reduction in the best case (last two rows of Table 1). It is noticable that the 3-gram FLM brings gain to the DBN as no gain was observed between 2 and 3-grams with the HMM system [6]. An explanation could be the use of a global (*generalized*) backoff procedure which provides better probability estimates for unobserved word sequences.

7. CONCLUSION

A multi-stage stochastic system combining generative (DBN) and discriminative (CRF) models for spoken language understanding has been proposed and evaluated. DBN models are used to derive sequences of basic concepts and values, then CRF models add the modalities and specifiers to the concepts. Taking benefit of a new segmental annotation scheme, this model has been applied successfully to the highly challenging MEDIA tourist information retrieval task (390 concept tags including hierarchical constituents and more than 2k concept values).

The DBN/CRF understanding system compares favorably with the reference HMM-based system in terms of understanding error rate. DBN alone have comparable performance and the impact of the CRF is clearly shown by a 3.3% improvement after its introduction to classify modality. A future extension of the system will be to derive automatically the optimal classes to be segregated in each stage. Also, DBN and FLM offer varieties of new features which remains to been explored (introduction of stems and POS in the FLM features, temporal constraints on the concept decoding...).

8. REFERENCES

[1] H. Bonneau-Maynard et al, "Results of the French MEDIA evaluation campaign for literal understanding", LREC, 2006.

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

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

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

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

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

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

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

[9] G. Ji and J. Bilmes, "Backoff Model Training using Partially Observed Data: Application to Dialog Act Tagging", HLT-NAACL, 2006

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

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

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

[13] Y.-Y. Wang and A. Acero, "Discriminative Models for Spoken Language Understanding", ISCA Interspeech, 2006.

[14] J. Lafferty et al, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", ICML, 2001.