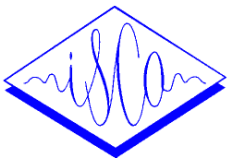

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

STOCHASTIC MODELING OF SEMANTIC CONTENT FOR USE IN A SPOKEN DIALOGUE SYSTEM

Magne H. Johnsen[1] *Trym Holter*[2] *Torbjørn Svendsen*[1] *Erik Harborg*[2]

[1]Norwegian University of Science and Technology (NTNU), N-7491 Trondheim, Norway
[2]SINTEF Telecom and Informatics, N-7465 Trondheim, Norway
mhj@tele.ntnu.no trym.holter@sintef.no

ABSTRACT

A key issue in a spoken dialogue system is the successful semantic interpretation of the output from the speech recognizer. Extracting the semantic concepts, i.e. the meaningful phrases, of an utterance is traditionally performed using rule based methods. In this paper we describe a statistical framework for modeling (and decoding) semantic concepts based on discrete hidden Markov models (DHMMs). Each semantic concept class is modeled as a multi-state DHMM, where the observations are the recognized words. The proposed decoding procedure is capable of parsing an utterance into a sequence of phrases, each belonging to a different concept class. The phrase sequence will correspond to a concept segmentation and class identification, whilst the semantic entities constituting each phrase contain the semantic value.

The algorithm has been tested on a dialogue system for bus route information in Norwegian. The results confirm the applicability of the procedure. Semantically relevant concepts in input inquiries could be identified with 6.9% error rate on the sentence level. The corresponding segmentation error rate was 8.6% when concept segmentation information was available during training. Without this information, i.e. if the training was performed in an embedded mode, the segmentation error rate increased to 23.5%.

1. INTRODUCTION

The speech understanding module plays a key role in every spoken dialogue system, as it should interpret the output of the speech recognition engine and provide a semantic description of the user input. The core of this procedure is to extract the meaningful phrases within an utterance. These phrases are usually termed the *semantic concepts*.

The speech understanding procedures traditionally belong to the family of rule based methods. The designed rule set should cover all syntactic and semantic cases, and significant expertise is thus required in the design. Statistical methods on the other hand rely heavily on large amounts of data, preferably dialogue samples collected from the domain for which the speech understanding module is designed. The main problem is then the acquisition and annotation of the data in terms of the *semantic concept classes*.

During the last decade, several statistical techniques have been reported. In both [1] and [2], a global HMM is designed, with each state corresponding to one unique concept class. Other solutions involve statistical decision trees [3] and probabilistic concept graphs [4]. The latter work combines a stochastic context free grammar and a concept class bigram.

In this paper, we investigate a procedure based on discrete hidden Markov models (DHMMs) [5]. The goal is to identify the semantic concepts, typically manifest as one or more short phrases or even single words, embedded in a spontaneously formed utterance. Each concept class is modeled by a DHMM, and the concept models are connected according to some kind of semantic language model.

A similar scheme is tested in [5], motivated by the success of embedded training of HMMs in speech recognition, i.e., under the hypothesis that it suffices to label each utterance in the training data by the corresponding *concept sequence*. However, this means that the correspondence between each word and the concept to which it belongs is ignored. We feel that the algorithm deserves a closer investigation, both because of its simplicity, but also because it seems like the scheme is well suited for integration with a speech recognizer in a spoken dialogue system. Thus, we have investigated the procedure with more carefully annotated data which allows training of the concept models in a bootstrap fashion, in order to reveal its full potential.

2. THE DHMM STOCHASTIC MODELING FRAMEWORK

The pattern recognition approach to semantic concept decoding is, given an utterance X, to select the concept sequence C for which the conditional probability is highest. That is, C' is the selected concept sequence if

$$C' = \underset{C}{\text{argmax}}\ P(C|X). \tag{1}$$

As usual, the training process does not easily permit characterization of these conditional probabilities, but rather generates an estimate of the conditional pdf $p(X|C)$. By Bayes rule, the relationship between these is given by

$$P(C|X) = \frac{p(X|C)P(C)}{p(X)}. \tag{2}$$

This equation implies that the semantic model should consist of two parts, the concept phrase model represented by the conditional pdf $p(X|C)$ and the concept language model represented by the probability $P(C)$.

We will evaluate the quality of the semantic model by its performance with respect to decoding of an unknown utterance into the

correct concept sequence. However, equally important is its ability to segment the utterance correctly, i.e., to assign each of the words in the sentence to the correct concept class.

2.1. The DHMM Concept Phrase Model

The HMM is very well known in the speech research community as a generating model, and as such it is well suited for modeling of the concept phrases. Each word is treated like a discrete symbol collected from a finite vocabulary. The state-specific discrete pdf thus determines the probability for each word to occur in the given state.

The use of multi-state DHMMs makes it possible to model the internal time structure of the concept phrases. It is thus clear that the topology of the HMMs will carry important information about the concepts, and it should be able to handle semantic phrases of varying lengths. In this work we have utilized left-to-right HMMs with one to four states. In each case, we have added skip-transitions between every state in a left-to-right manner. This way a multi-state HMM can be a valid model even for the shortest possible phrase, which is comprised by a single word.

2.2. The Concept Language Model

The a priori probability of a given length M concept sequence C, $P(C) = P(c_1, c_2, \ldots, c_M)$ can be written

$$P(C) = \left\{ \prod_{m=2}^{M} P(c_m | c_{m-1}, \ldots, c_1) \right\} P(c_1). \qquad (3)$$

For a given text corpus, it is difficult to estimate these conditional pdfs for a wide span, and it is usually assumed that c_m depends only on the preceding $n-1$ concepts. This leads to a standard N-gram concept language model. In this work we have investigated the performance of a bigram and compared it to an unigram (which neglects the history) and an 0-gram (which both neglects the history and assumes a uniform a priori concept probability).

2.3. Integration of the Semantic Network into a Dialogue System

A mixed initiative dialogue system often employs a speech recognition engine with a dialogue-state dependent language model. The reconfigurable language models are usually based on word bigrams or finite state word networks. The recognizer hypotheses are fed into the semantic decoder in a post-processing step.

As our decoder is purely statistical, we have the option of integrating the concept model into the speech recognizer, i.e., using our concept model directly as the recognizer language model. To do this, the concept model must be expanded into a DHMM-state network. The word probabilities in each concept-state pdf are then used for weighting of the acoustic scores obtained by the corresponding word HMMs. By estimating separate concept language models for each dialogue state (or for different groups of dialogue states), the language model seen by the speech recognizer will vary according to the current dialogue state. This line of thought is in accordance with an important principle stated in [4], i.e., to employ available sources of information as early as possible.

3. TASK DEFINITION

3.1. The Application

The idea of DHMM-based semantic modeling has been investigated in the domain of bus travel information for the city of Trondheim in Norway. The task was chosen due to several reasons. First of all, the dialogue has a manageable complexity. We also have access to the bus company's databases, and a text-based NLP inquiry system, BusTUC [6], already exists. BusTUC requests the users to formulate a complete inquiry in a single, preferably grammatically correct sentence, and it is available through the web. A spoken dialogue demonstrator called TABOR [7] employing a system driven dialogue has been developed on the basis of BusTUC. A pilot version is currently available through a public phone number. The total vocabulary of the speech recognizer is approximately 900 words. This includes some 700 place names in addition to bus numbers, hours and minutes, etc. We are currently developing a mixed initiative version of the bus travel information system, and we plan to test the DHMM semantic modeling framework as a part of this.

3.2. Training and Test Data

The web-based information system BusTUC has been operational for several years, and a significant amount of inquiry text data is thus available. In addition to this material, 100 real human-human dialogues from an operator service was recorded and annotated in an early stage of the TABOR project. A coarse semantic analysis of these databases showed that even though the spontaneously spoken utterances differ significantly from the written sentences, we found that most of the relevant semantic phrases were similar in the two cases. We thus believe that the text database is well suited for our experiments, although it is clear that a well-behaved semantic filler model will be of crucial importance in order to model the irrelevant and out-of-vocabulary (OOV) words which play a greater role in spontaneous speech than in written text.

We selected 3000 inquiries from the BusTUC-log for our application. The sentences were checked in order to ensure that each constituted a relevant question. The sentences were segmented and labeled using 32 carefully selected concepts, and later randomly divided into a training set (2000 inquiries) and a test set (1000 inquiries).

3.3. Semantic Concepts

The original annotation into 32 classes was based on the rule that each concept phrase should contain a single semantic entity, like e.g., BEFORE_TIME, AFTER_TIME, EXACT_TIME, etc. A closer analysis showed that the 32 concepts could be reconstructed from 13 broader concept classes. For instance, a single concept phrase like TIME would contain one or several semantic entities. The new set of 13 concept classes is shown with examples in table 1.

Note that the filler class contains in-vocabulary words like 'how' and 'travel', in addition to OOV words ('can', 'I') which are labeled by 'gar'. The in-vocabulary words in the filler class are words that contribute to the semantic phrases in one or several of the other concept classes.

Concept	*Example 1*	*Example 2*
BUS	bus twenty four	the airport bus
TICKET	ticket for children	ticket card for a month
BUS_GEN	bus connection	departure
DAY	today	on Sunday
N-S	directly	without change of bus
REL	next	the last one
PRICE	how much	the price
QUE	which bus	how do I
TIME	at sixteen thirty	between eight and ten
DUR	how long	travel time
FR_PL	from Ila	passing Bakli
TO_PL	towards town center	Lade
FIL	how gar gar travel	(how can I travel)

Table 1: Phrase concept classes with examples.

4. EXPERIMENTS

The concept models were trained using the HTK toolkit [8]. In order to investigate the importance of the segmentation borders, we trained concept phrase models both in bootstrap and embedded modes from the 2000 inquiries in the training data. In each case we experimented with 1-, 2-, 3- , and 4-state HMMs. The unigram and bigram concept language models were trained from the same data, and we also created a 0-gram for reference purposes. The bigram model included back-off transitions for concept pairs that were not observed in the training data.

Even if 2000 inquiries were estimated to be sufficient to capture the statistical structure of the phrases, it is far from enough to supply reliable statistics for the largest word classes. This is most evident for the place names (FR_PL and TO_PL), but is also observed for bus numbers (BUS) and points of time, i.e., hours (TIME) and minutes (TIME). Thus, for each of these four word classes, the accumulated state distribution probability mass was redistributed among all members of the word class according to a uniform probability assumption.

4.1. Results

The quality of the semantic models has been evaluated by its performance with respect to decoding of an unknown utterance into the correct concept sequence. This measure has been denoted *concept accuracy*, and is calculated both on the sentence and concept levels. In the following section we report the corresponding *concept error rates* (CER-snt and CER-cncpt). In these calculations, the filler concept class is ignored, as it does not contribute to the semantic contents of the utterance.

Equally important to the CER is the algorithm's ability to segment the utterance correctly, i.e., to assign each of the words to the correct concept class. This is quantified by the *word-tag accuracy*, i.e., the sequence of tags as assigned to each word by the algorithm is compared to the annotated test data. Again we report the corresponding error rates on the sentence (WTER-snt) and word levels (WTER-wrd). Note that the sentence level word-tag error rate by definition is at least as large as the corresponding concept error rate.

First of all, we investigated the influence of the different concept language models. The semantic concepts were modeled by 13 3-state DHMMs, trained in a bootstrap mode. The resulting models were tested with a 0-gram, a unigram, and a bigram. The results are shown in table 2, and points out the importance of the concept history, as the bigram clearly outperforms the two other language models.

Error type	*0-gram*	*1-gram*	*2-gram*
CER-snt	13.0%	12.5%	6.9%
CER-cncpt	3.8%	3.7%	2.0%
WTER-snt	14.3%	13.7%	8.6%
WTER-wrd	2.6%	2.5%	1.5%

Table 2: Error rates achieved with 3-state DHMMs trained in a bootstrap mode.

The main goal of the experiments has been to evaluate the general performance of the DHMM-framework, and in particular to compare embedded versus bootstrap training with respect to the resulting accuracy. In tables 3 and 4, we report error rates for HMMs trained in bootstrap and embedded modes, respectively. In both cases we have tested 1-, 2-, 3- , and 4-state HMMs, and the experiments have been performed with the bigram concept language model.

Error type	*1-state*	*2-state*	*3-state*	*4-state*
CER-snt	21.0%	7.3%	6.9%	8.9%
CER-cncpt	6.7%	2.2%	2.0%	2.5%
WTER-snt	73.4%	9.7%	8.6%	10.7%
WTER-wrd	11.4%	1.8%	1.5%	1.8%

Table 3: Error rates achieved with a bigram language model and DHMMs trained in a bootstrap mode.

Error type	*1-state*	*2-state*	*3-state*	*4-state*
CER-snt	24.1%	6.4%	7.2%	6.0%
CER-cncpt	7.8%	1.9%	2.1%	1.7%
WTER-snt	78.0%	15.5%	23.5%	23.6%
WTER-wrd	12.6%	2.6%	3.7%	3.7%

Table 4: Error rates achieved with a bigram language model and DHMMs trained in an embedded mode.

From tables 3 and 4, we first observe that the best results are achieved with the 3-state models in the bootstrap mode, and the 2-state models in the embedded mode. In both cases, the 1-state models are very poor compared to the others. This confirms that a multi-state HMM is crucial in order to capture the time structure within the concepts. We also find that the HMMs trained in the bootstrap mode clearly outperforms the embedded mode HMMs with respect to word-tag accuracy. The concept error rate is comparable for the two different schemes. This result is not surprising, given that the HMMs trained in embedded mode is not supplied with concept segmentation information during training.

Overall, the best results are achieved with the 3-state HMMs trained in bootstrap mode and a bigram concept language model. Note however that we do not claim that using 3-state HMMs for all

concepts is the optimal topology for modeling of our data. On the contrary, we have observed error rates comparable to the sentence level CER of 6.9% and WTER of 8.6% achieved in these experiments, with a combination of 1-state and 2-state HMMs with a more directed structure than used in these experiments. Thus, we believe that it will be possible to improve these results through a thorough HMM topology design adapted to each of the given concepts.

4.2. Error Analysis

We performed a detailed analysis of the sentences which contained errors for the 3-state HMMs trained in the bootstrap mode. It turned out that the word-tag errors could be split into three different groups. First of all, a large group of errors which do not influence the semantic content were discovered. This is exemplified by the inquiry "I am standing at Prinsen_kino, how do I get to Tyholt?". The automatically segmented utterance differ from the reference for only one word, i.e., 'at'. This word is not semantically relevant as long as the place name itself has been correctly tagged in the FR_PL concept class:

LAB:<FIL> gar gar gar at <FR_PL> Prinsen_kino <QUE> how
<FIL> gar gar gar <TO_PL> to Tyholt
REC:<FIL> gar gar gar <FR_PL> at Prinsen_kino <QUE> how
<FIL> gar gar gar <TO_PL> to Tyholt

The second kind of errors can be easily corrected in a simple post-processor stage. A typical error that is observed is that the decoder breaks time intervals into two consecutive TIME-phrases as in this example:

LAB: . . . <TIME> between sixteen and sixteen thirty. . .
REC: . . . <TIME> between sixteen and <TIME> sixteen thirty. . .

The final type of errors were unrepairable errors. Less than half of the errors were of this type. The following decoded sequence shows an example of this error type:

LAB:<REL> next <BUS> two <TO_PL> passing Ila
<FR_PL> from Tempe
REC:<REL> next two <TO_PL> passing Ila
<FR_PL> from Tempe

In this example, the user asks for the next bus number 2, while the system will decode this as an inquiry for the two next buses, irrespective of the bus number. We believe that many humans will interpret this question the same way as our system. In this particular case, the problem could be resolved through communication with the application back-end, by noting that bus number two actually does not pass the given bus stops.

5. CONCLUDING REMARKS AND FURTHER WORK

The results from the experiments show that the described DHMM-based semantic modeling framework is able to capture the relevant semantic structure in the given inquiries. In this application, the best results so far have been achieved with a bigram concept language model and a 3-state DHMM for each of the 13 concept classes. As expected, the HMMs trained in bootstrap mode outperformed the corresponding models trained in embedded mode, with respect to the ability to correctly segment utterances.

Future work in this direction will concentrate on implementing the proposed semantic language model in an integrated acoustic/semantic framework. Furthermore, dialogue state specific semantic networks will be developed in the mixed initiative system we are currently developing. Finally, we think that concept-value confidence measures based on the integrated acoustic/semantic modeling framework will be an interesting topic for further research.

6. ACKNOWLEDGMENTS

This work was financed by the Norwegian Research Council and Telenor R&D.

7. REFERENCES

[1] R. Pieraccini, E. Levin, and E. Vidal, "Learning how to understand language," in *Proc. European Conf. on Speech Commun. and Techn. (EUROSPEECH)*, (Berlin, Germany), pp. 1407–1412, Sept. 1993.

[2] W. Minker, S. Bennacef, and J.-L. Gauvain, "A stochastic case frame approach for natural language understanding," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, (Philadelphia, USA), pp. 1013–1016, Oct. 1996.

[3] R. Schwartz, S. Miller, D. Stallard, and J. Makhoul, "Hidden understanding models for statistical sentence understanding," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, (München, Germany), pp. 1479–1482, IEEE, Apr. 1997.

[4] B. Souvignier, A. Kellner, B. Rueber, H. Schramm, and F. Seide, "The thoughtful elephant: Strategies for spoken dialog systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 8, pp. 51–62, Jan. 2000.

[5] P. P. Boda, "From stochastic speech recognition to understanding: An HMM-based approach," in *Proc. of the 1997 IEEE Workshop on Speech Recognition and Understanding*, (Santa Barbara, USA), pp. 57–64, IEEE, Dec. 1997.

[6] T. Amble, "BusTUC–a natural language bus oracle," in *Applied Natural Language Processing Conference*, (Seattle, USA), Apr. 2000.

[7] M. H. Johnsen, T. Svendsen, T. Holter, and E. Harborg, "TABOR - a Norwegian spoken dialog system for bus travel information," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, (Beijing, China), Oct. 2000.

[8] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book V2.2*. Entropic Ltd., Jan. 1999.