Stochastic Spoken Natural Language Parsing in the Framework of the French MEDIA Evaluation Campaign

Dirk Bühler and Wolfgang Minker

University of Ulm
Deptartment of Information Technology
89081 Ulm/Donau, Germany
{dirk.buehler,wolfgang.minker}@uni-ulm.de

Abstract

A stochastic parsing component has been applied on a French spoken language dialogue corpus, recorded in the framework of the MEDIA evaluation campaign. Realized as an ergodic HMM using Viterbi decoding, the parser outputs the most likely semantic representation given a transcribed utterance as input. The semantic sequences used for training and testing have been derived from the semantic representations of the MEDIA corpus. The HMM parameters have been estimated given the word sequences along with their semantic representation. The performance score of the stochastic parser has been automatically determined using the MEDIAVAL tool applied to a held out reference corpus. Evaluation results will be presented in the paper.

1. Introduction

In this paper we report results for a stochastic parsing component applied on a large corpus of dialogues in French, recorded in the framework of the EVALDA-MEDIA evaluation campaign.

The EVALDA project[1] is financed by the French Ministry of Research in the context of the Technolangue programme. The aim of the project is to establish a reusable evaluation infrastructure for the language engineering sector in France and for the French language including organisation, logistics, language resources, evaluation protocols, methodologies and metrics.

Within EVALDA, the MEDIA task is the reservation of hotel rooms. Using a Wizard-of-Oz (WOZ) system that simulates a tourist information phone service, a total of 1,257 dialogues have been recorded from 250 different speakers where each caller carried out 5 different hotel reservation scenarios. The corpus containing 70 hours of dialogues has been transcribed by ELDA (Bonneau-Maynard *et. al.*, 2005). A context-independent semantic representation has been defined for that purpose. The aim of the MEDIA evaluation campaign is to test an automatic evaluation methodology for human-machine dialogue systems. The evaluation is based on a paradigm that uses test sets taken from a spoken language dialogue corpus, a semantic representation of the dialogue and common evaluation metrics. This protocol is designed to test the capacity of dialogue systems on a semantic level (cf. Figure 1), both taking into account and not taking into account, the context of the dialogue.

In order to validate the evaluation protocol and the semantic representations, an evaluation campaign has taken place where each partner in the project tests their system. The task chosen is hotel room reservation, with touristic information as an additional point of entry into the dialogue. In the remainder of this paper, the dialogue context is not taken into account.

The common semantic representation has been agreed up and formalized by the consortium. Each utterance (dialogue turn) is divided into semantic blocks. A block is annotated with a set of attributes including the *mode*, the underlying *concept* and the normalized form of the *semantic value*. Each semantic block spans individual words, such that all the words are covered by exactly one block. Additional information, such as noise events, can be annotated, but is discarded by our stochastic parser. An example of the MEDIA corpus structure is shown in Figure 2.

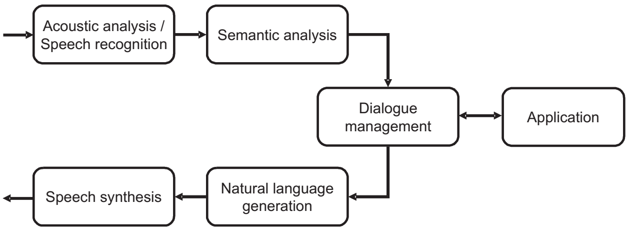
Acoustic analysis / Speech recognition
Semantic analysis
Dialogue management
Application
Speech synthesis
Natural language generation

Figure 1: Spoken Language Dialogue Systems architecture. The performance of the system is tested the semantic level, using the output of the semantic analysis module.

In the stochastic parsing techniques to be evaluated on the MEDIA task, statistical modeling techniques are used to parse the speech recognizer output into a semantic representation. The models are derived from the automatic analyses of large corpora of naturally-occurring utterances along with their semantic representations. Such stochastic methods have been applied, for instance, in BBN-HUM (Schwartz *et al.*, 1996), AT&T-CHRONUS (Levin *et al.*, 1998) and, more recently, by He and Young (2003, 2005).

The statistical models to parse the recognizer output transcripts into a semantic representation are derived from the user utterances and their corresponding semantic representations. The training data preparation and techniques to enhance the accuracy of the stochastic parser will be discussed here. Evaluation results with the official MEDIA test data will be provided.

The remainder of this paper is structured as follows. In Section 2 we review the fundamentals of HMM-based stochas-

[1] http://www.elda.org/

```
<Turn startTime="98.606" endTime="101.546" speaker="spk2" channel="telephone">
<SemDebut identifiant="55" mode="+" concept="null"/>
<Sync time="98.606"/>
est-ce qu'
<SemFin/>
<SemDebut identifiant="56" mode="+" concept="lienRef-coRef" valeur="pluriel"/>
ils
<SemFin/>
<SemDebut identifiant="57" mode="+" concept="null"/>
possedent
<SemFin/>
<SemDebut identifiant="58" mode="?" concept="hotel-services" valeur="salleMusc"
acte_dial="query"/> des salles de musculation
<SemFin/>
</Turn>
```

Figure 2: Example of the MEDIA corpus structure representing the semantic annotation of one single user utterance.

tic parsing, before continuing with the description of the mapping between the MEDIA corpus data and the HMM-based format in Section 3. In Section 4, we discuss optimization strategies for the stochastic model. A performance assessment based on the official `mediaval` tool is presented in Section 5. Finally, Section 6 concludes our paper with a summary and an outlook to future directions.

2. Stochastically-based Semantic Analysis

In the stochastic parsing approach, the semantic decoding consists of maximizing the conditional probability $P(S|O)$ of a state (semantic) sequence $S = (s_1, s_2, \ldots, s_n)$ given the observation (word) sequence $O = (o_1, o_2, \ldots, o_n)$ over all possible state sequences S (Rabiner and Juang, 1986). Like in (Beuschel *et al.*, 2004), the semantic sequences used for training and testing have been derived from the semantic representations of the MEDIA corpus.

Reformulating $P(S|O)$ by using Bayes' rule, we obtain for the desired state sequence:

$$[S]_{opt} = \arg\max_S \{P(S)P(O|S)\} \tag{1}$$

With a first-order HMM, statistical independence between non-adjacent states is assumed and the problem is simplified to the computation of

$$\begin{aligned} [S]_{opt} &\approx \arg\max_S P(S|O, \text{HMM}) & (2)\\ &= \arg\max_S P(O|S, \text{HMM}) \cdot P(S|\text{HMM}) & (3) \end{aligned}$$

with

$$P(S|\text{HMM}) = P(s_1|\text{init}) \cdot P(s_2|s_1) \ldots P(s_T|s_{T-1}) \tag{4}$$
$$P(O|S, \text{HMM}) = P(o_1|s_1) \cdot P(o_2|s_2) \ldots P(o_T|s_T) \tag{5}$$

Unlike Eqn. (1), each conditional probability in the products in Eqn. (4) and Eqn. (5) depends only on one state, thereby significantly reducing the computational complexity.

The maximization problem in Eqn. (3) may be resolved by estimating the HMM parameters. These include the initial state distribution probabilities $P(s_i|\text{init})$, the bigram state transitions probabilities $P(s_j|s_i)$ and the observation symbol probabilities $P(o_m|s_j)$. The most convenient way to calculate the parameters is to simply count the events in the training corpus, e.g. to calculate $P(s_j|s_i)$ by summing up all occurring state transitions from s_i to s_j and dividing this amount by the number of state transitions from s_i to any state (Maximum Likelihood estimation). However, the test data may contain observations that never occurred during training. To adequately estimate rare and unseen events, a back-off technique (Katz, 1987) has been applied for parameter estimation using the CMU-Toolkit (Rosenfeld, 1995).

Throughout decoding, the Viterbi algorithm allows to determine the most likely sequence of semantic labels (corresponding to the model states) given a word (or observation) sequence and the HMM. This algorithm can be visualized by a trellis where all possible states s_i are plotted against the observation sequence. The nodes in the trellis can then be specified by $n_{i,t}$, so that for each discrete point in time t, there exists a node for each possible state s_i. Using an ergodic HMM, each node at time t is connected to each node at time $t-1$. Viterbi decoding is processed from left to right in the trellis.

3. Training Data Preparation

The semantic MEDIA representation (cf. Figure 2) is not in a form that may be directly used by the model parameter estimator of the stochastic parsing component. As mentioned earlier, the representation is based on semantic blocks, delimited by `<SemDebut>` and `<SemFin>` tags in the XML notation. The `<SemDebut>` tag contains the `mode`, `concept`, and `valeur` attributes that are valid for the words before the next `<SemFin>` tag. As can be seen in Figure 2, some words may be annotated with a "null" concept, i.e. these words were ignored in the construction of the semantic content of the utterance. In addition, the words "des salles de musculation" are assigned a mode value of "?", which indicates that the content of corresponding block is meant as a question.

In our stochastic parsing approach, the MEDIA representation is automatically been converted into sequences of semantic labels, one label for each word. An example of the converted MEDIA corpus structure is shown in Figure 3.

Each word in the user utterance corresponds to a semantic

```
UTT: <s> est-ce qu' ils possedent des salles
         de musculation </s>
SEM: <s> [+][null][*0]
         [+][null][*1]
         [+][lienRef-coRef][pluriel][*0]
         [+][null][*0]
         [?][hotel-services][+][*0]
         [?][hotel-services][+][*1]
         [?][hotel-services][+][*2]
         [?][hotel-services][+][*3] </s>
```

Figure 3: Example of the converted MEDIA corpus structure used for the stochastically-based semantic parsing. Each word in the user utterance *do they have a fitness room* yields a corresponding semantic label.

label extracted from the MEDIA corpus. Semantic blocks yield identical yet subsequently numbered concepts. For the HMM modeling, the block structure has to be mapped to a word sequence structure, consisting of observation-label pairs. To this end, we define for each word a *feature vector*, consisting of features that represent local (within the block) and global (within the utterance) information.

Feature	*Description*
m	mode: the mode attribute of the current block
c	concept: similarly, the concept attribute
v	value: similarly, the valeur attribute
q	question: flag indicating if this utterance contains a block with mode="?"
n	negative-mode: similarly, for mode="-"
r	reservation: similarly, a concept with "-reservation" suffix
h	hotel: for "-hotel" suffix
z	chambre: for "-chambre" suffix

Table 1: Summary of local and global features defined for each word of an utterance.

The list of features is described in Table 1. The features m, c, and v are calculated on the basis of the current block, whereas q, n, r, h, and z are identical for the whole utterance. The intuition behind such global features is to provide a form of utterance-wide context for the interpretation.

The "valeur" attribute can not be treated directly as a feature, since this attribute may have a closed class of values (such as "hotel services"), or an open class (such as hotel or city names) depending on the concept. A direct usage would therefore be problematic for reliable statistical modeling. Instead, the approach taken in our experiment is to try to identify the words in the semantic block that correspond to the "valeur" for the open value case. This is relatively straightforward, such as in city names or numbers, which can be directly used as values. In the same manner, month names can be mapped to the number representation used in the "valeur" attribute (JANVIER → 01 etc.) Some cases require more attention, e.g. leading prepositions and determiners are mostly removed from the value. Therefore, we define the feature value "+" for words that should be integrated in the normalized result, and "-" for words that should be ignored.

In order to normalize the observation words, apostrophes and French special characters are replaced by similar (capitalized) ASCII variants.[2] Also, filler words like "EUH" are removed.

Word classes (cf. Table 2) provide a means of treating different but similar observations in the same manner. For instance, the words "lundi" (Monday) and "mardi" (Tuesday) will most likely have the same function in a scheduling corpus, and so they should be treated identically by the HMM model. This can be achieved by replacing these words by their word class representative, and calculating the statistics on this basis. In this way, word classes also help to make the statistics more robust and reliable, since the observation counts for each member of the word class are merged into the count of the word class representative. Finally, word classes can be a way of dealing with known words that happen not to occur in the training data, especially for large classes such as city names. A summary of the word classes used in this experiment is given in Table 2.

Word class	*Corresponding Words*
/NUM/	ordinary numbers: DEUX, TROIS, QUATRE...
/ORD/	ordinal numbers: DEUXIEME, TROISIEME, ...
/ORDS/	ordinal numbers with plural ending: DEUXIEMES, TROISIEMES, ...
/MONTH/	month names: JANVIER, FEVRIER, ...
/PLACE/	place names: AGEN, AIX-EN-PROVENCE, ...
/HOTEL/	hotel names: ABENNA, ACACIAS, ...

Table 2: Word classes for the MEDIA corpus.

The stochastic parser model has been trained using 9,344 utterance transcriptions along with their semantic label sequences. The dialogues result in a relatively large lexicon size (about 1,900 different words). The average utterance length is about eight words per utterance. Applying word classes allows to reduce the lexicon size to 1,284 distinct words.

We have defined 288 basic semantic units, representing the different values the semantic attributes (mode, concept, and semantic value) can assume. These units combine to 711 distinct states.

4. Stochastic Parsing Strategies

Oriented Model In the example utterance (Figure 3) some labels, such as [?][hotel-services][+], are repeated for successive words. In an oriented model topology, submodels are defined within the ergodic HMM. For example, the label corresponding to the word *des* becomes [?][hotel-services][+][*0] and the one corresponding to *salles* are associated to [?][hotel-services][+][*1]. Thus, [?][hotel-services][+] is replaced by a sub-model corresponding to subsequently numbered states resulting in a left-to-right propagation within the HMM. Such an oriented model

[2] With the exception of the preposition "à": this word is marked with a flag in order not to be confused with verb "a".

topology allows to significantly improve the performance of the parser whilst significantly increasing the number of states.

Observation Context Model The difficulty in natural language understanding relies in the fact that words may yield different meanings depending on the contiguous words. First order HMMs can only model dependencies between adjacent states. To account for these effects, *context* information is introduced.

We define *contextual observations* as the current word along with a pre-determined number of its adjacent words and call an observation without any adjacent words a *non-contextual observation*.

The number of contextual observations increases significantly if the context is extended to more than one word. Best decoding results have been obtained experimentally by defining the current word along with the two adjacent words on the left and on the right, i.e. five words in total.

5. Performance Assessment

Due to time constraints, the stochastic spoken natural language parser has not been evaluated within the framework of the MEDIA campaign. However, automatic evaluations have been performed on the official test corpus using the MEDIA semantic evaluation paradigm. The transcribed test data consists of 25,115 words in 3,003 utterances (about 8.4 words per utterance.)

The performance score was determined with the MEDIAVAL tool. The stochastic parser achieved an overall score of 71.4% of correctly labelled semantic blocks. If the semantic attributes are evaluated individually, the scores for mode, concept and value sequences at the semantic block level are 86.5%, 80.2% and 82.4%, respectively.

An analysis of the errors indicates that a lot of problems with the normalized attribute values, such as place and hotel names exists. Also, in date expressions, implicit context dependencies seem to be used in deriving a normalized date value that contains a month number, even though this may not be present in the semantic block.

Besides these specific cases of errors, some errors that are characteristic to the HMM approach can be noted. Firstly, single-word “null” blocks are often appended to either the preceding or the succeeding block. This is most likely due to the “penalty” block transitions incur in the stochastic model. Similarly, the parser has problems reproducing the strict annotation scheme for incomplete phrases or repetitions. These spontaneous speech effects are usually labelled as “null” blocks, since these words do not contribute relevant information to the interpretation of the utterance. The stochastic parser, however, will regard some incomplete phrases as “almost perfect” and will assign them in most cases the label that the complete phrase would have.

6. Summary and Outlook

A number of improvements and optimizations are likely to be achieved within the HMM framework. Firstly, grammatical information should be used/(taken into account) to a larger degree. By and large, the human-annotated corpus consists of grammatical phrases, at least on a chunk level. So, a separate part of speech tagging module might contribute valuable information (as additional features).

On the other hand, French is a language with a rich morphology, especially regarding the conjugated verb forms of different tenses. As the stochastic model ignores these relations and similarities, a lemmatization should help to build a more compact and therefore more reliable model.

We are currently investigating the use of the Probabilistic Context Free Grammar (PCFG)-based Stanford parser for the MEDIA task. The interesting feature of the Stanford parser is its ability to use a factored stochastic model, consisting of a pure PCFG and a word dependency grammar. The main challenge consists of mapping the semantic block structure into a structure that is adequate for the Stanford parser. Although the Stanford parser has already been applied to different languages (including different models of parts of speech and syntactic categories), these applications constitute syntactic approaches whereas the MEDIA annotation scheme is a mostly semantic one.

7. References

Beuschel C., Minker W., and Bühler D. (2004), “Strategies for Optimizing a Stochastic Spoken Natural Language Parser,” ICSLP.

Bonneau-Maynard H., Rosset S., Ayache C., Kuhn A., and Mostefa D. (2005), “Semantic annotation of the French Media dialog corpus,” Eurospeech.

He Y. and Young S. (2005), “Semantic Processing using the Hidden Vector State Model,” Computer Speech and Language, 19(1): 85-106.

He Y. and Young S. (2003), “Hidden Vector State Model for Hierarchical Semantic Parsing,” ICSLP.

Katz S.M. (1987), “Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer,” IEEE Transactions on Acoustics, Speech and Signal Processing, 35(3):400–401.

Levin E., Pieraccini R., and Eckert W. (1998),”Using Markov decision process for learning dialogue strategies,” ICASSP.

Schwartz R., Miller S., Stallard D., and Makhoul J. (1996), “Language Understanding Using Hidden Understanding Models,” ICSLP.

Rabiner L.R. and Juang B.H. (1986), “An introduction to Hidden Markov Models,” IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 3(1).

Rosenfeld R. (1995), “The CMU Statistical Language Modeling Toolkit, and its use in the 1994 ARPA CSR Evaluation,” ARPA Workshop on Spoken Language Technology.