Stochastically-Based Natural Language Understanding Across Tasks and Languages

Wolfgang Minker

Spoken Language Processing Group
LIMSI-CNRS
91403 Orsay cedex, FRANCE
email: minker@limsi.fr
http://www.limsi.fr/TLP

ABSTRACT

A stochastically-based method for natural language understanding has been ported from the American ATIS (Air Travel Information Services) to the French MASK (Multimodal-Multimedia Automated Service Kiosk) task. The porting was carried out by designing and annotating a corpus of semantic representations via a semi-automatic iterative labeling. The study shows that domain and language porting is rather flexible, since it is sufficient to train the system on data sets specific to the application and language. A limiting factor of the current implementation is the quality of the semantic representation and the use of query preprocessing strategies which strongly suffer from human influence. The performances of the stochastically-based and a rule-based method are compared on both tasks.

1. INTRODUCTION

In this paper, we report on our experience in porting a stochastically-based natural language understanding component across tasks and languages. Stochastically-based methods have been applied in the BBN-HUM [5], the AT&T-CHRONUS [7] systems and at LIMSI-CNRS [1] for the American ARPA-ATIS (Air Travel Information Services) task[1]. Since the stochastically-based decoding techniques are rather similar across the sites, the systems differ primarily in the definition of the knowledge sources, which are represented in the form of semantic labels. In a stochastically-based method, correspondencies between these labels and the corresponding words are automatically learned from a large annotated training corpus and memorized in the form of model parameters. These parameters are then used by the semantic decoder to generate the most likely semantic sequence given an unknown input query.

Another travel-related application is explored in the context of the ESPRIT Project 9075 MASK (Multimodal-Multimedia Automated Service Kiosk). A spoken language system in French has been developed at LIMSI [3] for this task, which allows users to obtain train travel information including schedules, services and fares. We investigate language and domain portability by porting the stochastically-based semantic analyzer presented in [1] from the American ATIS task to the French MASK application. For the American ATIS task, a rule-based version [2, 4] was used to automatically produce a corpus of semantic representations for training the stochastically-based component in [1], enabling a direct comparison between both methods. In MASK, we have focused on creating the corpus of semantic annotations via an iterative semi-automatic labeling approach. The similarity of both tasks enables us to apply equivalent query preprocessing and semantic decoding strategies. The preprocessing includes a lexical analysis and category unification which reduces redundancies in the corpus. We also model the observations in context in order to improve the reliability of the decoding.

[1] Systems developed within the framework of ATIS allow the user to acquire information derived from the Official Airline Guide about fares and flight schedules available between a restricted set of cities within the United States and Canada

2. KNOWLEDGE SOURCES

The parameters of the stochastic model are estimated given sequences of preprocessed words (observations) and their corresponding semantic labels (states).

2.1. Semantic representation

For the understanding components we use a semantic case grammar to represent the meaning of the spoken request [4]. This formalism is considered to be more suitable for spontaneous speech, than a grammar based on a purely syntactic analysis, typically performed by context-free grammars.

Applying the case grammar to the specific task consists of defining the meaningful concepts and the corresponding reference words used to identify the concepts. The MASK concepts determined by analysis of queries taken from the training corpora are train-time, fare, connection, type, book, service, reduction and train-type. Associated to each concept is a set of constraints which are introduced by semantic markers.

Je souhaiterais ***réserver*** *une place pour le tarif le moins cher*

<book>
(c:num-seat) ↦ une
(c:fare-comparative) ↦ le-moins-cher

Figure 1: *Frame-based semantic case grammar representation for MASK exemplified for "je souhaiterais réserver une place pour le tarif le moins cher (I would like to book a seat with the least expensive fare)".*

For the example in Figure 1, the concept <book> is identified by the reference word ***réserver***. The marker *place* (m:num-seat) designates *une* to be a constraint on the number of seats (c:num-seat). The semantic markers do not appear in the frame in Figure 1. Most of the concepts and constraints for ATIS are found in the train travel domain, albeit with slightly different significations. For example, the constraints related to arrival and departure times can be mapped directly, whereas the concept type corresponding to the aircraft type in ATIS corresponds to the type of train in MASK (TGV[2], EuroCity, etc.).

In the probabilistic framework, a sequential representation of semantic labels is obtained by aligning the concepts, markers and constraints (third column in Figure 2). In the frame-based representation of the case grammar in Figure 1, the semantic annotation is not exhaustive. Only words related to the concepts and its constraints are considered. In order to label the training data, the additional semantic label *(null)* is associated with those words, that

[2] TGV = Train à Grande Vitesse (High Speed Train)

are judged not to be useful by the case frame analyzer for the specific application, e.g. *je, souhaiterais, pour* and *le* in the example.

je	{*filler*}	*(null)*
souhaiterais	{*filler*}	*(null)*
réserver	{*réservation*}	*<book>*
une	{*1*}	*(c:num-seat)*
place	{*place*}	*(m:num-seat)*
pour	{*filler*}	*(null)*
le	{*filler*}	*(null)*
tarif	{*prix*}	*(m:fare-comparative)*
le-moins-cher	{*prix-minimum*}	*(c:fare-comparative)*

Figure 2: *Example query "je souhaiterais réserver une place pour le tarif le moins cher", its preprocessed form and the corresponding semantic labels in a sequential representation.*

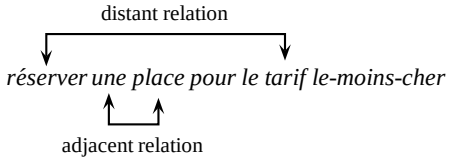
2.2. Query preprocessing

The case grammar is an economic semantic representation which ignores a substantial number of words that are not significant for the semantic decoding. Many inflected forms are also attributed to the same semantic categories. This redundancy increases the model size unnecessarily and makes parameter estimation and decoding less reliable. A query preprocessing, similar to that used in the ATIS system [1], removes redundancies and introduces additional contextual information.

Lexical analysis The first step is lexically-based using a look-up table. Inflected words are replaced with their corresponding base forms and semantically-related words are clustered. Non-relevant or out-of-domain words are assigned to a {filler} category (second column in Figure 2). Even though yielding an important simplification, this data manipulation has an important drawback as the isolated lexical entries are judged without accounting for their context. There is thus a risk of incorrectly clustering words which are ambiguous out of context. In an attempt to make the lexical analysis less arbitrary, the semantic function of the word has been introduced as an additional parameter. Only those words and synonyms covering identical semantic functions are clustered into identical categories. For example, *je réserve le* ***premier*** *en* ***première*** *classe non-fumeur* (*I book the first in the first class no smoking*) is preprocessed into {filler} {réservation} {filler} **premier** {filler} **première** classe {non-fumeur}, instead of {filler} {réservation} {filler} {**1**} {filler} {**1**} classe {non-fumeur}. Taken in context, **premier** and **première** can be respectively understood in the sense of *first train* and *first class* without any surrounding marker. This example also illustrates the difficulty of obtaining a robust grammar representation containing distinctive marker-constraint relations: instead of the {filler} unit, an explicit marker function could be assigned to the prepositions *le, en*. However, *le* cannot be used as a marker in the context of the example query in Figure 2.

Category unification In the domain of information retrieval, a large number of lexical entries correspond to database values, which can sometimes be clustered. Observed in the training data, 19 task-related categories have been defined for ATIS, including airport names, flight identifiers, etc., and 8 for the MASK application due to a comparatively smaller domain coverage, e.g. station names, train types, etc..

Contextual observations The lexical analysis and category unification reduce the redundancies in the input queries and thus the model size. This in turn allows us to define the more detailed contextual observations. The example query



contains both, adjacent relations, e.g. between *une* and *place* and longer distance ones, such as between *réserver* and *tarif*. The adjacent relations can be unambiguously decoded by a bigram language model, e.g. (c:num-seat) (m:num-seat) in Figure 2. However, if only context-independent observations are considered, the system fails on this example, because it identifies two concepts, *réserver* $\mapsto$ <book> and *tarif* $\mapsto$ <fare>, even though the request clearly is for a reservation, and in this case, *tarif* $\mapsto$ (m:fare-comparative). The meaning of any given word is contextual. In the human understanding process, the semantic significance of a word is properly determined in context of the current query or even the following or preceeding queries. The simplest stochastic implementation aligns semantic labels providing surface forms rather than deep semantic structures and considers the observations on an isolated word-by-word basis. In order to improve the accuracy of the model, we introduce relevant concept-related semantic information in the form of contextual observations. A look-up table, established from the training corpus, first associates to each isolated word in the input query the corresponding non-contextual local concept, which signifies for the example:

réserver	*une*	*place*	*pour*	*le*	*tarif*	*le-moins-cher*
⇓	⇓	⇓	⇓	⇓	⇓	⇓
[book]	[empty]	[empty]	[empty]	[empty]	[fare]	[empty]

réserver, réservation, for instance, are reference words for the concept <book>, therefore assigned along with all the words that may trigger this concept to the local concept [book]. *tarif, prix* are associated to [fare]. Since *une, place, etc.* are never used as reference words, they are assigned to an [empty] concept. The context-dependent decoding then uses this local information and defines the word in its left (-) and/or right (+) context of local concepts, which for the same example query yields:

réserver	*une*	*place*	*pour*	*le*	*tarif*	*le-moins-cher*
[fare]$^+$	[fare]$^+$	[fare]$^+$	[fare]$^+$	[fare]$^+$	[empty]$^+$	[empty]$^+$
[empty]$^-$	[book]$^-$	[book]$^-$	[book]$^-$	[book]$^-$	[book]$^-$	[book]$^-$

tarif in the left context of the local concept [book]$^-$ is less likely to be decoded as a reference word than *réserver* given the [fare]$^+$ as a right context, thus *réserver* $\mapsto$ <book> and *tarif* $\mapsto$ (m:fare-comparative).

3. SEMANTIC DECODER

The semantic decoding consists of maximizing the conditional probability $P(s_1^T|o_1^T)$ of some state sequence s_1^T given the observation sequence o_1^T. Using Bayes rule, this probability is reformulated and the following optimality criterion is defined:

$$[s_1^T]_{opt} = \underset{s_1^T}{\arg\max} \left\{ P(s_1^T) P(o_1^T|s_1^T) \right\} \quad (1)$$

Given the dimensionality of the sequence o_1^T, the estimation of the likelihood $P(o_1^T|s_1^T)$ is replaced by estimating the parameters of a Hidden Markov Model (HMM). We use an ergodic model topology which allows all states to follow each other. The model parameters consist of bigram state transitions probabilities $P(s_j|s_i)$

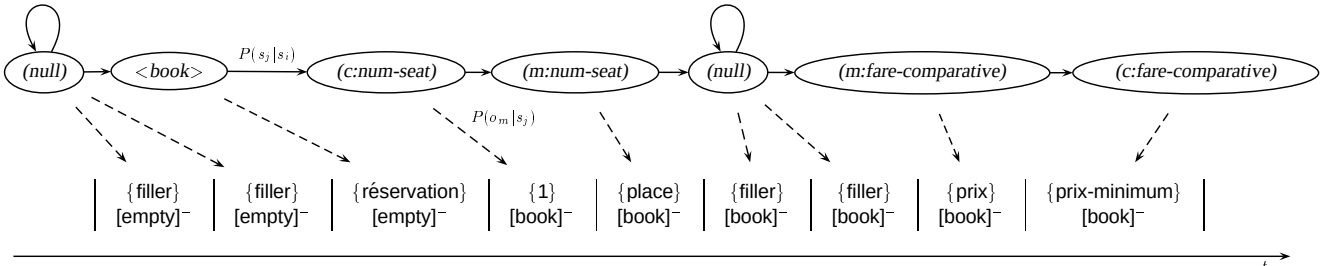


Figure 3: *Semantic decoding progresses on a path through the model. It generates left-contextual observation sequences of the preprocessed example query "je souhaite réserver une place pour le tarif le moins cher".*

and the observation symbol probability distribution $P(o_m|s_j)$ in state j, which are estimated using a back-off technique [6]. In Figure 3, the particular path through the Markov Model is shown for the example query. The progression through the state sequence of semantic labels generates a sequence of observation vectors each of them containing a preprocessed entry along with left-contextual local concepts. This temporal progression and sequence generation is guided by the state transition and observation probabilities, previously learned from a large number of correspondencies between states and observation vectors in the training data.

For contextual observations, the corresponding probability distribution is:

$$P(o_m|s_j) = P(l_m, c_m|s_j) \quad (2)$$

where l_m is the preprocessed word and c_m the observation context. In practice, better performance results from more reliable estimates obtained by interpolation of contextual and non contextual observation models:

$$P(o_m|s_j) = \lambda P(l_m, c_m|s_j) + (1-\lambda)P(l_m|s_j) \quad (3)$$

The parameter λ has been experimentally determined. For ATIS, a left and right context model is applied with $\lambda = 0.9$, for MASK, a left context model with $\lambda = 0.3$.

4. CORPUS ESTABLISHMENT

The ATIS and MASK data were collected by subjects solving predefined travel scenarios. The speech is spontaneous and unconstrained. Using the MASK spoken language system [3, 8], over 25,000 queries have been recorded. The ATIS corpus results from a multi-site data collection effort and consists of approximately 13,000 speech queries.

4.1. Semi-automatic data annotation

In order to estimate the model parameters, the stochastic method requires semantically annotated corpora. For ATIS, the annotations were automatically produced by the rule-based component in [2] and tailored to this system, but were suboptimal for the stochastic method [1]. In MASK, the semantic representation was determined independently from the already existing rule-based case system in [3]. If no previous knowledge other than the grammar formalism is introduced, the semantic corpus can be better adapted to the stochastic component. The optimum semantic representation for a specific application is not previously defined, but is determined throughout the development process.

A semi-automatic, iterative approach was used to annotate the MASK data, as illustrated in Figure 4. The MASK training corpus of 10,500 queries was divided into four subsets containing 500, 1,000, 3,000 and 6,000 queries. Parses were manually determined for the first 500 sentences (*initialization*). The *model* parameters were estimated on this initial subset. Then the iterative procedure started: using the model, the decoder annotated each query in the following subset. Even with a small amount of training, the vast majority of the parses were correct, and thus very little effort was required to correct them manually. For the data correction, each semantic label of the sequence had to be verified. Typical errors at this stage of development issued from an increase in domain coverage when annotating the new data. The annotated sets were merged and the model parameters re-calculated for further query annotation. These steps were iterated until the complete training set was semantically annotated and corrected. Annotation became faster as more data were available for improved parameter estimation. This automatic approach simplified system development considerably and enabled us to accomplish the porting within a period of 15 working days.

An important issue with such a semi-automatic technique is to assure the consistency of the semantic representations in the different data subsets. Even though the rule-based system in [2] was iteratively modified as a function of new recorded training data, we used a snapshot version of this system to produce a homogeneous semantic corpus for ATIS [1]. Semantic representations for MASK were subject to constant modifications since they were adapted and tuned throughout the iterative annotation. The quality of the corpus was monitored through periodic tests on the training data which revealed inconsistencies and weaknesses at each stage of the development.

4.2. Data characteristics

The data characteristics for the French MASK and American ATIS training corpora are compared in Table 1. Roughly the same number of queries have been used for both applications, in order to create equivalent training conditions. The main reason for the comparable lexicon sizes (1,449 versus 1,577) is that the French language is highly inflected than English and provides a variety of words with identical root forms, e.g. *réservant, réserverai, réserves, réservez, réserve, réservé, réservée* are various inflections of the word *réserver* (*to book*). As a result of the query preprocessing, the initial lexicon size has been considerably reduced for both applications: a total of 162 preprocessed entries for MASK and 299 for ATIS are used. Applying a left-contextual observation model for MASK and a left and right context model for ATIS result in the respectively 2,284 and 8,885 observation vectors, applied along with the semantic labels to estimate the model parameters. For MASK 74 semantic labels have been determined throughout the data annotation, compared to 112 for ATIS.

		MASK	ATIS
#queries		10,500	10,718
lexicon size	*raw data*	1,449	1,577
	+ lexical analysis	394	955
	+ category unification	162	299
	+ context	2,284	8,885
#semantic labels		74	112

Table 1: *Comparing data characteristics of the MASK and ATIS training corpora for statistical modeling in natural language understanding.*

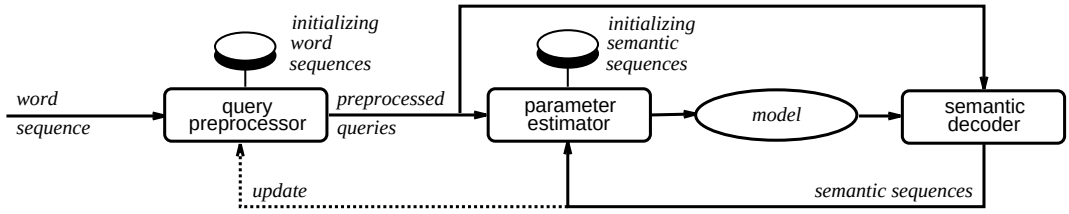


Figure 4: *Semi-automatic procedure to establish a corpus of semantic labels for MASK and to create and update the preprocessing components.*

5. EVALUATION

Performance evaluations have been carried out on the stochastically-based and rule-based natural language understanding components for MASK and ATIS (Table 2) in order to validate the porting. The MASK understanding components were tested using 15 travel scenarios containing 726 queries. The ATIS test data consisted of the 445 type A queries from the official ARPA-ATIS December 1994 Benchmark test. The performance was assessed at the semantic sequence level, comparing the concepts and constraints with previously defined reference labels. The components were also evaluated on the accuracy of the mere system responses returned to the user, which are the retrieved database responses in ATIS and the natural language response generated in MASK. The evaluation methodology is described in more detail in [1, 3].

	Semantic sequence (%)		*Response (%)*	
	STOCHASTIC	RULE-BASED	STOCHASTIC	RULE-BASED
MASK	7.2	13.8	8.3	9.4
ATIS	13.7	14.4	18.7	16.9

Table 2: *Semantic sequence and response error rate (%) for MASK and ATIS comparing stochastically- and rule-based components.*

At the semantic sequence level, the stochastic clearly outperforms on the rule-based implementation in MASK [3] obtaining 7.2 % compared to a 13.8 % error rate. Unlike for ATIS, where the difference in performance is small for the two methods the independent design of the stochastic system in MASK does not limit its performance by eventual shortcomings of the rule-based method. The stochastic implementation also profits from the mutual information between all the semantic labels. If an explicit marker is incorrectly decoded or does not exist, the surrounding words yield the function of implicit markers. This makes a decoding of the associated constraints more robust, such as in *réserver une place*, where the reference word *réserver* implicitely introduces the constraint *une*. However, the system frequently fails to identify long distant marker-constraint relations. For the rule-based MASK system, 68 % of the errors involve concept identification due to an incorrect triggering of reference words or the identification of multiple semantic concepts. This is mainly due to the difficulty the rule-based decoding strategy has in coping with conflicting slots.

A priori we may expect that the response evaluation should yield the highest performance, as even an incorrect semantic representation can potentially yield a correct system reaction. However, this is only true for the rule-based implementation in MASK. The stochastically-based understanding component was not integrated in the framework of the spoken language system, where the semantic representation is also oriented so as to be able to respond appropriately to the user. The current decoder outputs the meaning of an isolated query regardless of the ongoing dialog context. The reponse errors made by the stochastic component are therefore due to ignoring dialog specificities, which need to be addressed when the component is fully integrated in an end-to-end system. For ATIS, the performance loss for both implementations is attributed to the difficulty of matching the response generation to the *min-max* reference answer strategy adopted by the ARPA community.

6. CONCLUSION

In this paper we have described the porting of a stochastically-based natural language understanding component using a semantic case grammar from the American ATIS task to the French MASK application. The design of the stochastic component focalized on the creation of a corpus of semantic labels, obtained through semi-automatic data annotation. This technique simplified the system development considerably and consistency checks assured a homogeneous semantic corpus. The porting was accomplished within a rather short time frame. By adapting the semantic labels to the method, the performance of the stochastic component in MASK is able to outperform the rule-based method.

The study shows, that domain and language porting of a stochastic method is rather flexible, since instead of translating and adapting the rule-based case grammar, it is sufficient to train the components on the application and language specific data sets. Another advantage of the sequential semantic representation is the progression of mutual information, increasing the robustness of the decoding. The lexical query preprocessing reduces the model size, but has the disadvantage of requiring human intervention. Another shortcoming is the rather flat semantic representation which does not allow the modeling of nested structures, whose necessity was illustrated by the intrusion of contextual observations. Instead of simply aligning the semantic labels on a one-level word-by-word basis, a hierarchical, e.g. tree-structured, representation would be more appropriate. Experiments using such a more powerful semantic representation are in progress.

Acknowledgement The author acknowledges the contribution of Sophie Rosset to the comparative evaluations of the natural language understanding components.

7. REFERENCES

[1] W. Minker, S. Bennacef, and J. L. Gauvain. A Stochastic Case Frame Approach for Natural Language Understanding. *Proc. ICSLP-96*.

[2] W. Minker and S. Bennacef. Compréhension et Évaluation dans le Domaine ATIS. In *Proc. Journées d'Études en Parole, JEP-96*.

[3] J. L. Gauvain, S. Bennacef, L. Devillers, L. Lamel, and S. Rosset. Spoken Language Component of the MASK Kiosk. In K. Varghese and S. Pfleger, editors, *Human Comfort & Security of Information Systems*, pages 93-103. Springer, 1997.

[4] S. Bennacef, H. Bonneau-Maynard, J. L. Gauvain, L. F. Lamel, and W. Minker. A Spoken Language System For Information Retrieval. In *Proc. ICSLP-94*.

[5] R. Schwartz, S. Miller, D. Stallard, and J. Makhoul. Language Understanding Using Hidden Understanding Models. In *Proc. ICSLP-96*.

[6] S. M. Katz. 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, 1987.

[7] E. Levin and R. Pieraccini. CHRONUS - The Next Generation. In *Proc. ARPA/HLT-95*.

[8] A. Life, I. Salter, J. Temem, F. Bernard, S. Rosset, S. Bennacef, and L. Lamel. Data Collection for the MASK Kiosk: WOz vs Prototype System. In *Proc. ICSLP-96*.