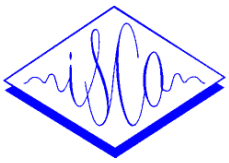

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

Automatic Learning in Spoken Language Understanding

Roberto Pieraccini Zor Gorelov Esther Levin Evelyne Tzoukermann

AT&T Bell Laboratories
600 Mountain Avenue, Murray Hill, NJ 07974, USA

ABSTRACT

In this paper we propose a mechanism for learning the parameters of a model that constitutes the basis of the natural language component of a speech understanding system. The model defines a representation of the meaning of a sentence as a sequence of elemental semantic units. The sentence production mechanism, in this paradigm, is equivalent to a noisy channel whose input is the sequence of meaning units and whose output is a sequence of acoustic observations. The decoding (i.e., the understanding) is then formalized as the problem of finding the meaning given the acoustic representation. The automatic estimation of the model parameters is possible if a statistically significant set of sentence examples is available, and if each sentence is provided with the correct meaning. Unfortunately a database of sentences annotated with their meaning is not available at the moment. Instead we have a database, within the DARPA ATIS project [4], in which each sentence is given a correct answer. In this paper we discuss the problem of automating the training procedure and we give some experimental results.

10.21437/ICSLP.1992-84

1 Introduction

In this paper we refer to a natural language (speech or text) understanding system as a system that, given as input a suitable representation of a sentence (e.g., a sequence of acoustic measurements for spoken utterances or simply a sequence of word identifiers for written sentences), will produce as output a representation of the sentence meaning, suitable for performing the action required by the subject that uttered the sentences. The output of the language understanding system, i.e., the representation of meaning, is then converted into the required action, which could be the generation of the proper answer or the execution of a command.

Looking back at the history of speech recognition, we find that a rule-based paradigm was proposed as an alternative to the template-matching and stochastic approaches. In the rule-based approach to speech recognition, a human expert encodes the rules that allow the identification of basic units of speech (phonemes, allophones, phones, syllables, words, etc.) based on the observation of complex features extracted from the speech signal. These rules are then integrated with lexical and syntactic rules according to the application language. An *inference engine* analyzes the features extracted from the incoming speech and tries to determine the spoken words, according to the specified rules. On the other hand, the statistical approach (often referred to as the *brute force* approach) does not generally rely on complex features nor on a priori human knowledge. In the statistical approach to speech recognition, a stochastic model of each basic unit is designed, and its parameters are estimated from a number of examples of the chosen speech units. The only human supervision required in the training (i.e., the estimation of the model parameters) is the labeling of each example as a sequence of the basic recognition units (i.e., if the units are words and if the examples are sentences, each utterance must be annotated with the exact sequence of spoken words). For these reasons, during the last decade, when computers became able to deal with large amounts of data in reasonable time, the stochastic approach to speech recognition outperformed the rule-based approach, so much that the literature today does not report any rule-based speech recognizer. We can conclude that the high performance of today systems is achieved by the ability to use large databases of examples.

We believe that a similar situation will occur in the discipline of language understanding. Current natural language understanding systems are generally based on rules that are generally designed by an expert. This procedure makes maintenance, updating, and generalization of a system to other tasks a very expensive and difficult operation. Moreover, the set of rules that a human expert can think of for describing a language will hardly be exhaustive because of the variety of expressions present in that language. The situation is even more critical when spoken language is taken into account. Spoken language, often ungrammatical and idiomatic, generally follows rules that are different from written language rules. Moreover, in spoken language, there are phenomena like false starts and broken sentences that do not appear in written language. Many examples of the idiosyncrasies of spoken natural language can be found by analyzing a recently collected database of dialogues [10] in the airline reservation domain, within the DARPA ATIS project [4]. An extreme example of this kind of sentence is the following:

> *FROM uh sss FROM THE PHILADELPHIA AIRPORT um AT ooh THE AIRLINE IS UNITED AIRLINES AND IT IS FLIGHT NUMBER ONE NINETY FOUR ONCE THAT ONE LANDS I NEED GROUND TRANSPORTATION TO uh BROAD STREET IN PHILELD PHILADELPHIA WHAT CAN YOU ARRANGE FOR THAT* [1]

It is clear from this example that, although the spoken language often does not follow the grammatical rules of English, it is still able to convey meaning. This suggests that an understanding system designed for spoken language should allow enough flexibility to deal with ungrammatical expressions and disfluencies, and should incorporate a mechanism that permits learning new expressions from examples. Furthermore, since we are interested in developing a speech understanding system, the understanding model should define a framework that allows an easy and natural integration with the speech recognizer.

Following the above considerations we designed a framework, called *CHRONUS* [2], based on a stochastic representation of conceptual entities, that has the following features:

[1] This sentence was cited by Victor Zue, MIT, during the 5th DARPA Workshop on Speech and Natural Language, Harriman, NY, Dec. 1991.

[2] *CHRONUS* stands for Conceptual Hidden Representation of Natural Unconstrained Speech. This acronym is used both for the proposed representation paradigm as well as for the whole understanding system.

SHOW ME THE FLIGHTS TO BOSTON	(question,display) (subject,flight) (destin,BBOS)
HOW MUCH IS THE PRICE OF THE FLIGHT FROM ATLANTA	(question,display) (subject,fare) (destin,MATL)

Table 1: Example of keyword/pair representations of simple phrases within the ATIS domain.

- The model parameters can be learned from examples of sentences during a training stage. The examples must be annotated with the exact sequence of conceptual entities.
- A bootstrapping procedure can be developed that allows us to use data that is not originally annotated in terms of the defined conceptual entities, with the minimum amount of human supervision.
- The understanding system can be naturally integrated with a speech recognizer [8].

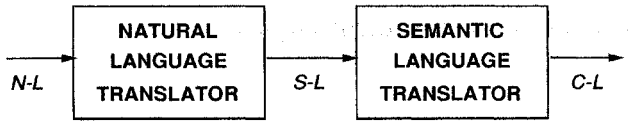

Figure 1: Understanding as a cascade of two language translators

2 Formalization of the Language Understanding Problem

In this section we propose a formalization of the spoken language understanding problem in terms of the noisy channel paradigm. According to this paradigm we think of an utterance as a corrupted and noisy form of the symbolic representation of its meaning. Here we assume that the meaning of a sentence can be expressed by a sequence of basic units $\mathbf{M} = m_1, m_2, \ldots, m_{N_M}$ and that there is a sequential correspondence between each m_j and a subsequence of the acoustic observation $\mathbf{A} = a_1, a_2 \ldots a_{N_A}$. This hypothesis, although very restrictive, was successfully introduced also in [7]. According to this model of the spoken sentence production, one can think of decoding the original sequence of meaning units directly from the acoustic observation. The decoding process can be based on the maximization of the a posteriori probability $P(\mathbf{M} \mid \mathbf{A})$ (*maximum a posteriori probability* decoding, or MAP decoding).

The problem now consists in defining a suitable representation of the meaning of a sentence in terms of basic units. The representation we chose was inspired by the *semantic network* [1] paradigm, where the meaning of a sentence can be represented as a relational graph whose nodes belong to some category of concepts and whose arcs represent relations between concepts or linguistic cases. In our representation, each unit of meaning consists of a *keyword/value* pair $m_j = (k_j, v_j)$, where k_j, is a conceptual relation (referred to as *concept* hereafter), (e.g., *origin, destination, meal* in the ATIS domain), and v_j is the value with which k_j is instantiated in the actual sentence (*value* hereafter). (e.g., *Boston, San Francisco, breakfast*). Given a certain application domain, we can define the concept dictionary $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_{N_k}\}$, and for each concept γ_j we can define the set of possible values it assumes $\Upsilon^j = \{v_1^j, v_2^j, \ldots, v_{N_j^j}\}$. Examples of meaning representation for phrases in the ATIS domain are given in Table 1. Details of the particular implementation of this paradigm for the ATIS task can be found in [5, 6, 8, 9].

3 Training of the Understanding System

It is interesting to refer to the understanding system as a cascade of two language translators, as shown in Fig. 1. The *natural language translator* analyzes the input sentence expressed in natural language N-L and transforms it into the representation of its meaning expressed in a semantic language S-L (in our system the semantic language coincides with sequences of keyword/value pairs, as in the examples of Table 1). The *semantic language translator* transforms the semantic language S-L into code expressed in computer language performing the action required by the sentence (in our system C-L is the language for accessing the database of flights). Training the understanding system for a new application consists of the following steps:

1. Defining a semantic language S-L that can express the whole variety of meanings of the application.
2. Designing or updating the *semantic language translator* according to the newly defined semantic language.
3. Collecting a large set of sentences.
4. Providing each collected sentence with its correct meaning described in the defined semantic language. Thus each token in the training set is a pair (n_i, s_i), where n_i is a sentence in N-L and s_i a sentence in S-L.
5. Training the *natural language translator*.

The first three steps require a great deal of human intervention and we do not see, at the moment, any way of automating the procedures. However, some procedures can be designed for reducing human supervision in the last two steps, namely the semantic annotation and the estimation of the model parameters. Since the solution of these problems is not completely general, but depends on the particular application the system is designed for, we will refer in the following to the DARPA ATIS task.

3.1 Semantic annotation of the training set

ATIS stands for Air Travel Information System. This task [4] was built around a subset of the OAG (Official Airline Guide) database that includes only 10 American cities. Speech is being collected by different sites [10] giving each subject a scenario with a travel planning problem to solve. The subjects carry out a dialogue with a machine (through a human *wizard*) in order to solve the problem. The partial and final responses are presented to the subjects via a display or a speech synthesizer. The sentences uttered by the subject are recorded, transcribed, and annotated carefully. Among other information related to each recording session, the annotators include the following:

- A detailed transcription of the sentence in terms of spoken words, including hesitations, false starts, and some kinds of noise.
- A categorization of the kind of query. The query can be context-independent (class A: it can be answered regardless of the previous query); context-dependent (class D: it can be answered only with reference to a previous question); unanswerable (class X: it cannot be answered in the domain of the specified application).

HOW	*MUCH*	*IS*	*THE*	*PRICE*	*OF*	*THE*	*FLIGHT*	*FROM*	*ATLANTA*
question			subject					origin	

Table 2: Example of conceptual segmentation

- A minimal and a maximal reference answer. The answer to each valid sentence is represented either by a set of data extracted from the database, by a numerical quantity, or a by boolean value (yes or no). Deciding on the right answer to a question often depends on the interpretation of the sentence itself. A set of *principles of interpretation* [10] was compiled by a special committee within the DARPA ATIS project, and it is updated any time a new interpretation criterion is needed.

A correct answer should contain all the information included in the minimal answer, but no more than that included in the maximal answer. The hypothesized and reference answers are compared according to a methodology explained in [3], by a program called the *comparator*.

Therefore the annotation already included in the database does not directly map onto a semantic language that could serve as a meaning representation language. Besides, the current annotation in terms of the actual answer to the query cannot be transformed into a meaning representation since this is a one-to-many relationship, i.e., the same answer corresponds to a large number of different queries. The methodology we developed consists in establishing a training loop in which the output of the comparator is used as a feedback signal. The procedure is iterative. An initial model is used for translating each sentence into the corresponding meaning expressed with the defined S-L. Then the semantic language translator is used for generating the answer to each sentence, and finally the answer is compared with the reference answers. If the answer is correct, we assume that the representation of its meaning expressed in S-L is correct and it will be used in the next iteration for estimating a new model. The whole procedure can be summarized with the following steps:

1. Start with a reasonable model.
2. Generate an answer for each sentence in the new training set.
3. Compare each answer with the corresponding reference answer.
4. Use the meaning representation of the sentences that were given a correct answer to reestimate the model parameters.
5. Update the model and go to step 2

A certain number of sentences will still produce a wrong answer after several iterations of the training loop. The conceptual segmentation of these sentences may then be corrected by hand and included in the training set for a final reestimation of the model parameters.

3.2 Estimation of the model parameters

According to [5] the CHRONUS model consists of two sets of parameters, namely the *concept conditional bigrams* $P(w_i \mid w_{i-1}, c_i)$ and the *concept transition probabilities* $P(c_i \mid c_{i-1})$, where w_i is the $i-th$ word in the sentence and c_i is the concept it expresses. Both sets of parameters can be estimated starting from sentences that are segmented in terms of conceptual entities. An example of segmentation is shown in Table 2.

Set	Number of Sentences	Description
A	532	handlabled
B	446	annotated
C	195	annotated (oct-91)

Table 3: Description of the data sets used in the training experiment

Training set	% correct on set B	% correct on set C
A	48.2	63.5
A+smooth		72.3
A+T(B)	50.9	72.8
A+T(B)+smooth		73.3

Table 4: Results using the training loop described in the text. T(B) is the subset of B that was correctly answered by the system.

One way to provide an initial estimate of a model is by segmenting a set of sentences by hand. The initial model can be used for performing a forced segmentation of each sentence, and the so obtained segmentation can be used again for reestimating a new model, and so on, in the same fashion of the *segmental k-means* training algorithm [2]. However, in a preliminary experiment we didn't notice appreciable changes in performance iterating after the first segmentation.

Table 3 shows the sets of data used for testing the effectiveness of the training loop. All sentences are class A (context-independent) and belong to the MADCOW database. The conceptual segmentation of the sentences in set A was performed by hand; sets B and C consisted of the officially annotated sentences (set C corresponded to the October 91 test set). The results of this experiment are reported in Table 4. The first line in the table shows the results (as the percentage of correctly answered sentences) both on set B and on set C when the initial model, trained on the 532 hand-labeled sentences, was used. The second line shows the results on set C when the probabilities of the initial model were smoothed using the supervised smoothing technique described in [6]. The third line reports the accuracy (on both set B and set C) when the sentences that were correctly answered out of set B were added to the training set (this set is called T(B)) and their conceptual labeling was used along with set A for reestimating the model. It is interesting to notice that the performance on set C is higher than that obtained with supervised smoothing. The last line of Table 4 shows that supervised smoothing increases the performance by a very small percentage. The results of this experiment show that the use of automatically produced conceptual segmentation along with the feedback introduced by the comparator improves the performance of the system by an amount that is comparable with that obtained by a supervised procedure.

4 Conclusions

In this paper we reviewed the framework of understanding as a decoding process in presence of a noisy channel. We also discussed the implications arising in the implementation of a training procedure that will allow for the minimum amount of human supervi-

sion. Since in the current DARPA ATIS database only the correct answer of each sentence is given rather than its meaning, we introduced the *comparator* in the training loop. We show that using this procedure without hand-labeling any new training material we can actually improve the performance of a system that was initially trained on a set of manually annotated sentences.

References

[1] Simmons, R. *Semantic networks: their computation and use for understanding English sentences*, In Schank and Colby, eds Computer Models of Thought and Language, Freeman: San Francisco, 1973.

[2] Rabiner, L. R., Wilpon J. G., Juang B. H., "A segmental k-means training procedure for connected word recognition based on whole word reference patterns," *AT&T Technical Journal*, vol. 65 no. 3, pp 21-31, May/June 1986

[3] Bates, M., Boisen, S., Makhoul, J., "Developing an Evaluation Methodology for Spoken Language Systems," *Proc. of 3rd DARPA Workshop on Speech and Natural Language*, pp. 102-108, Hidden Valley (PA), June 1990.

[4] Price, P. J., "Evaluation of Spoken Language Systems: the ATIS Domain," *Proc. of 3rd DARPA Workshop on Speech and Natural Language*, pp. 91-95, Hidden Valley (PA), June 1990.

[5] Pieraccini, R., Levin, E., Lee, C. H., "Stochastic Representation of Conceptual Structure in the ATIS Task," *Proc. of 4th DARPA Workshop on Speech and Natural Language*, Asilomar (CA), February 1991.

[6] Pieraccini, R., Levin, E., "Stochastic Representation of Semantic Structure for Speech Understanding," *Proc. of EUROSPEECH 91,* Genova, Italy, September 1991.

[7] Prieto, N., Vidal, E., "Learning Language Models through the ECGI Method," *Proc. of EUROSPEECH 91,* Genova, Italy, September 1991.

[8] Pieraccini, R., Tzoukermann, E., Gorelov, Z., Gauvain, J. L., Levin, E., Lee, C. H., Wilpon, J. G., "A Speech Understanding System Based on Statistical Representation of Semantics," *Proc. of ICASSP 92,* San Francisco, CA, March 1992.

[9] Pieraccini, R., Tzoukermann, E., Gorelov, Z., Levin, E., Lee, C. H., Gauvain, J. L., "Progress Report on the Chronus System: ATIS Benchmark Results," *Proc. of Fifth Darpa Workshop on Speech and Natural Language*, Harriman, NY, Feb 1992.

[10] *MADCOW*, "Multi-Site Data Collection for a Spoken Language Corpus,", *Proc. of Fifth Darpa Workshop on Speech and Natural Language*, Harriman, NY, Feb 1992.