FROM F. DEINZER ET AL., *EUROSPEECH'99*, BUDAPEST, 1999, PP. 1987–1990.

LEARNING OF DOMAIN DEPENDENT KNOWLEDGE IN SEMANTIC NETWORKS

F. Deinzer, J. Fischer, U. Ahlrichs, E. Nöth

Chair for Pattern Recognition
University of Erlangen-Nuremberg
Martensstraße 3
D-91058 Erlangen, Germany
e-mail: {deinzer,fischerj,ahlrichs,noeth}@informatik.uni-erlangen.de
www: http://www5.informatik.uni-erlangen.de

ABSTRACT

For an efficient linguistic analysis of spoken queries a lot of domain specific knowledge is needed and usually has to be entered manually into the knowledge base of each domain. This makes the adaption of dialogue systems which base on explicit knowledge representation to new domains a very costly procedure. We use a frequency based statistical method combined with general hidden markov models in order to learn domain specific knowledge within a semantic network formalism. As a framework we use a dialogue system for German train timetable information. By means of experiments we show that our statistical approach is not only able to reach, but even outperforms previous results with manually entered restrictions.

1. INTRODUCTION

Speech understanding systems usually make use of explicit linguistic knowledge representation, for example, by means of rules and parsing algorithms or a semantic network formalism. For an efficient linguistic analysis of spoken queries in the context of a specific application domain, a lot of domain specific knowledge is needed and is usually entered manually into the corresponding knowledge base. This is a very time consuming procedure which contradicts the demand for flexibility for such systems. Thus, it seems to be obvious to make use of statistical methods (provided that a corpus of training data is available) in order to learn the domain specific knowledge quickly and without major costs. During the last few years, corpus-based approaches are emerging more and more, combined with knowledge-based techniques, in order to automate the fitting of a grammar to a certain domain [1, 2].

In our approach we learn domain specific linguistic restrictions in a semantic network. The semantic network formalism which serves as a framework allows a clear knowledge representation and an easy and uniform integration of speech with other sources of information (e.g. images). The control algorithm we use for knowledge processing during the linguistic analysis is based on iterative optimization and a bottom-up processing of a fine-grained task graph. This graph is automatically compiled off-line out of the knowledge base. knowledge which is represented as a fine-grained task graph. This approach enables an easy integration of statistical methods and provides the system with any-time and real-time capabilities.

2. KNOWLEDGE REPRESENTATION AND PROCESSING

We use a parallel version of the dialogue system EVAR [3] which answers queries about the German train timetable as a framework for our approach. The linguistic knowledge base of EVAR represents knowledge on 5 *levels of abstraction*: The *word-hypotheses* level is the most concrete one and represents the interface between speech recognition and speech understanding; on *syntax* level syntactic constituents are represented; the *semantic* level is used to model verb and noun frames with their deep cases for task independent interpretation; on the *pragmatic* level, semantic information is interpreted in the context of the application domain; The *dialogue* level is the most abstract one and models information about dialogue acts and sequences thereof.

The frame-based semantic network formalism ERNEST [4] we use represents knowledge about general terms, events, etc. in *concepts* C (e.g. a concept SY_NOUN represents knowledge about a noun on syntax level); the actual occurrence of a concept C in the sensor data is represented by an *instance* $I(C)$ (e.g. the noun "train" is represented by an instance $I(\text{SY\_NOUN})$). Relations between the concepts are established by *part-*, *concrete-*, and *specialization-links*.

The main components of a concept C are, besides its *parts* and *concretes*, its *attributes*, *structural relations* and a *confidence measure*, which computes the degree of confidence of $I(C)$ and its expected contribution to the success of the analysis. Each attribute, relation and the confidence measure references a *function* which computes its value when an instance for the corresponding concept is computed. Since there may be different possibilities for the actual realizations of a concept (e.g., a noun phrase can be composed by a proper noun: "Berlin", or a noun, an article, and an adjective: "the next train") *modalities* can be specified for the concepts. Each modality defines a combination of parts of the concept, which can realize a valid instance of it. This keeps the knowledge base compact, but in turn, ambiguities arise.

The task the linguistic analysis has to solve is to find an interpretation of the sensor data, given the knowledge base and the initial segmentation (here: word-hypotheses). Because of multiple occurrences of words (or word categories) in a word-chain and ambiguities in the knowledge base (arising from the modalities) there may be several competing interpretations. Thus, one uses a control-algorithm to search for the *optimal interpretation* (which corresponds to the best scored instance of that concept on the highest level of abstraction representing the goal of the analysis).

Our control algorithm treats the search for an optimal interpretation as a *combinatorial optimization* problem and solves it by means of *iterative* optimization methods, e.g. stochastic relaxation, simulated annealing, and genetic algorithms. For this purpose, a state of analysis (or interpretation) vector is introduced, which makes the assignment of exactly one modality to each ambiguous concept, and one initial segment (here: a word) to each concept on word-hypotheses level. Each specific value allocation of this vector reflects exactly one out of the possible interpretations. All possible value allocations of this vector build the *search space* of the analysis. Once a specific value allocation (i.e. an actual state of analysis) has been chosen, the corresponding interpretation can be computed clearly without ambiguities in a bottom-up way. In order to make this computation more efficient and to allow an efficient exploitation of parallelism, the concept-centered and well-structured knowledge base is, automatically and off-line, compiled into a fine-grained task graph (the so-called *attribute network*). It explicitly represents the dependencies of all attributes, relations, and confidence measures which have to be computed for an interpretation. The attributes, relations, and confidence measures are represented by the nodes, the dependencies between them by the directed links of the task graph. Nodes without predecessors (*primitive nodes*) represent attributes that provide the interface to the word-hypotheses and nodes without successors represent the confidence measure of the concept which represents the goal of the analysis.

Now, in each iteration step of the iterative optimization a current state of analysis is chosen and mapped onto the attribute network. The corresponding interpretation and its confidence value are then computed by a bottom-up processing of all nodes of the attribute network. Iteration steps are performed until the 'best' interpretation is found or the specified processing time for analysis has been used up. Notice that, when interpretation is stopped due to time limitations, a (suboptimal) interpretation is still available if at least one iteration was performed. At the moment, one iteration needs approximately 0.2 seconds of processing time.

3. LINGUISTIC RESTRICTIONS

In order to allow an efficient linguistic analysis, the enormous search space has to be reduced. This is done by making use of specific linguistic knowledge about syntax, semantics and pragmatics in the way of *linguistic restrictions*. These restrict the range of values for the attributes A^j (e.g. case, gender, syntactic class, semantic class, pragmatic class, etc.) of concepts from all possible values $A_i^j,\ i = 1, \ldots, N^j$ to a few ones. For example, the attribute gender of a word can be restricted from *masculine, feminine, neuter* to only *feminine*.

These restrictions are usually entered manually into the respective attribute-slots of the concepts in the knowledge base. The attribute network of our approach allows us to propagate top-down (from dialogue level to the primitive nodes on the word-hypotheses level) those restrictions concerning the properties of the initial segments (i.e. words) once before the analysis starts. This enables the rejection of word-hypotheses during analysis which could potentially be assigned to a primitive node, but which violate at least one of the restrictions at this node.

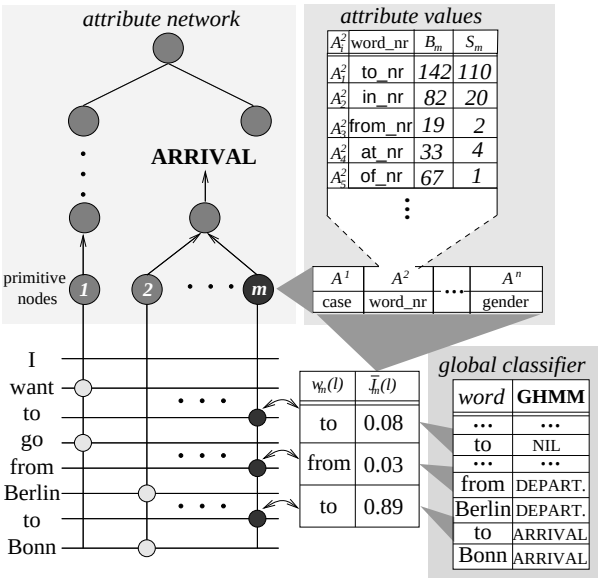

Figure 1. Principles of the confidence values and their sources of information.

The reduction of search space is thus achieved by a drastic reduction of the number of competing word-hypotheses at each primitive node. This led to the motivation for our approach: To learn linguistic restrictions at each primitive node within the attribute network by adding a learning component to the control of the analysis.

4. LEARNING OF DOMAIN DEPENDENT KNOWLEDGE

Together with the manually modelled linguistic restrictions a binary decision measure was used (i.e., either the restrictions are fulfilled by a word-hypothesis or not). The use of statistics will replace these manually entered restrictions and expand the semantic network by a new continuous measure component. This measure enables the automatic adaptation (both, off- and online) of parts of the linguistic knowledge to new application domains and the learning of domain specific occurrences of colloquialisms.

For this purpose we introduce a *rating function* which automatically learns the linguistic restrictions for each primitive node of the attribute network. During analysis, it computes at each primitive node k a rating for each word-hypothesis $w_k(l),\ l = 1, \ldots, L$ (number of word-hypotheses which can potentially be assigned to node k), which bases on the learned linguistic restrictions and on contextual information. This rating replaces the former binary decision and represents a measure the compatibility of a word-hypothesis with a primitive node. It can be further improved online after each new analyzed utterance.

The ratings are used as follows to speed up the analysis: They enable an *initialization* of the state of analysis vector by assigning an "optimal" word-hypothesis to each primitive node (the initialization of the state of analysis vector concerning the modalities of the concepts is described in [5]). Furthermore, they allow a *weighted* change of word-hypotheses during the optimization, when a new current state of analysis is chosen before a new iteration step.

Figure 1 illustrates the principle of the learned ratings. In this figure $k = 1, \ldots, m$ and node m has

the role of a preposition[1] (on pragmatic level, this primitive node belongs to a *place of arrival*). Thus, all words of the word lattice which are prepositions ("to", "from", and "to") can potentially be assigned to node m.
The learned linguistic restrictions for the attribute values (case, word_nr, ..., gender) at node m are illustrated in the upper right of Figure 1. Together with the results of the global classifier (lower right) which delivers coarse contextual information by means of the best word-chain, a rating for each preposition is computed: "to" = 0.08, "from" = 0.03, "to" = 0.89. As expected, the second "to" has the highest rating, since it is in fact the preposition of the place of arrival "to Bonn". Without the contextual information, the ratings would be 0.89, 0.33, and 0.89, respectively.

Learning of linguistic restrictions. The learning of the linguistic restrictions for a primitive node k is based on the *observations* B_k and the *successes* S_k. These correspond to the counted frequencies at node k of observed and successful attribute values $A_i^j(w_k(l))$ of word-hypotheses $w_k(l)$ for all sample utterances during the training. The observed attribute values belong to those word-hypotheses which could potentially be assigned to a primitive node (in Figure 1, for example, the attribute values case, word_nr, ..., gender of $w_m(1)$ = "to", $w_m(2)$ = "from", and $w_m(3)$ = "to"). The successful attribute values belong to those word-hypotheses which finally led to the optimal interpretation (in Figure1, $w_m(3)$). The training data consists of a set of utterances and their correct interpretation and is used by the learning algorithm to decide after each iteration step, if the optimal interpretation has already been found, i.e. if a correct assignment of word-hypotheses to primitive nodes is already available. Notice that the learning happens within the attribute network itself.
The frequencies $B_k(\cdot)$ and $S_k(\cdot)$ are used as the basis for the *restriction rating* $J_k(\cdot)$ in form of a maximized — as one word-hypothesis can have, for example, several cases — *success rate*

$$R_k^j(l) = \max_i \left(\frac{S_k(A_i^j(w_k(l)))}{B_k(A_i^j(w_k(l)))} \right) . \qquad (1)$$

It describes the chance of a word-hypothesis $w_k(l)$ with regard to its attribute value $A_i^j(w_k(l))$ to lead to a correct interpretation. In Figure 1 the success rate of "to" regarding the value of its attribute word_nr is $R_m^{\text{word\_nr}}(3) = S_m(\text{to\_nr}) / B_m(\text{to\_nr}) = 110/142 = 0.77$.
Since a specific attribute may contain more important information than another attribute at a specific primitive node (e.g., the attribute pragmatic_class is more important to find the noun of a place of arrival than the attribute word_nr) weights G_k^j are introduced for each attribute A^j at each node k. Thus, a weighted restriction rating is finally computed:

$$J_k(l) = \sum_{j=1}^{n} G_k^j \cdot R_k^j(l) \qquad \text{with} \sum_{j=1}^{n} G_k^j = 1 . \qquad (2)$$

[1]Since we have an explicit representation of all paths from the primitive nodes to the node representing the goal of analysis, we can automatically determine beforehand to which concepts on syntax, semantic, pragmatic, and dialogue level a primitive node is connected to.

The weights G_k^j are also learned automatically, based on the idea that if the success rates of all values of one attribute at a specific primitive node k are quite similar, we assume that this attribute is not relevant for the calculation of $J_k(l)$. As a measure of similarity we use the standard deviations σ_k^j over all N_k^j previously observed and successful attribute values A_i^j:

$$\sigma_k^j = \frac{1}{N_k^j} \sqrt{\sum_{i=1}^{N_k^j} \left(\frac{S_k(A_i^j)}{B_k(A_i^j)} - \frac{1}{N_k^j} \sum_{i=1}^{N_k^j} \frac{S_k(A_i^j)}{B_k(A_i^j)} \right)^2} . \qquad (3)$$

We use a gradient descent method for adapting the weights $(G_k^1, \ldots, G_k^n)^T$ with $(\sigma_k^1, \ldots, \sigma_k^n)^T$ representing the descent vector.

Computing the rating function. An improved analysis that takes advantage of the learned attribute values has to take into account, that — because our success rate is based exclusively on the counting frequencies $B_k(\cdot)$ and $S_k(\cdot)$ — $R_k^j(l) = 0$ for attribute values which were not observed in the training data but occur in the test data. To avoid this we use a smoothing technique [6, 7] which redistributes the frequencies for attribute values (in Equation 4 denoted as $a = A_i^j(w_k(l))$) in favour of missing or rare observations. The smoothing is shown for $S_k(\cdot)$ in Equation 4 and applied to $B_k(\cdot)$ accordingly. The smoothed frequencies $S_k^*(\cdot)$ and $B_k^*(\cdot)$ lead to the *smoothed success rate* $R_k^{*j}(l)$ and the *smoothed restriction rating* $J_k^*(l)$.

$$S_k^*(a) = S_k(a) + \frac{1}{2} \cdot \sqrt{1 + \sum\nolimits_n S_k(A_n^j)} . \qquad (4)$$

It further has to be considered, as already mentioned, that the context within an utterance is important for the analysis. Taking a look at competing word-hypotheses one notices that without contextual knowledge they will be judged equally in several cases. Consider, for example, the utterance in Figure 1: "I want to go from Berlin to Bonn". Without considering the prepositions in the context of "Berlin" and "Bonn" it is not possible to decide whether "Berlin" or "Bonn" is the place of departure. Thus, the confidence values of both nouns will be similar. Notice that the same case applies to the two "to"'s appearing in the utterance, as explained before. To reduce the control algorithms costs of optimization caused by word-hypotheses with similar $J_k^*(l)$, we introduce a *global classifier* that extends the restriction rating values with information about how well each word fits at a specific primitive node considering the context of the whole utterance. Therefore, each word-hypothesis $w_k(l)$ of the utterance $\boldsymbol{w}$ is labeled with its most probable pragmatic intention. This labeling is realized with generalized hidden-markov-models GHMM$(w_k(l)|\boldsymbol{w})$ [8], e.g.:

$$\underbrace{\text{I want to go}}_{\textbf{NIL}} \quad \underbrace{\text{from Berlin}}_{\textbf{DEPART.}} \quad \underbrace{\text{to Bonn.}}_{\textbf{ARRIVAL}}$$

Since we want the GHMMs to perform only a coarse labeling, six pragmatic labels representing cities of departure (DEPART.), cities of arrival (ARRIVAL), general cities (CITY), dates (DATE), times (TIME) and dispensable information (NIL) are sufficient for

our needs. The GHMM's classification result is combined with the smoothed restriction rating, finally leading to the *context-sensitive rating function* $\bar{J}_k(l)$: As already mentioned, all paths from the primitive nodes to the node representing the goal of analysis are explicitly represented in the attribute network. Thus, one can automatically assign a mark $\mathcal{B}(k)$ to each primitive node k. This mark denotes the pragmatic concept which can be reached on the path from the primitive node to the goal node. Now, each word-hypothesis $w_k(l)$ whose pragmatic label $\mathrm{GHMM}(w_k(l)|\boldsymbol{w})$ does not match the mark $\mathcal{B}(k)$ is devaluated by a *global factor* F:

$$\bar{J}_k(l) = \begin{cases} J_k^*(l) & : \ \mathrm{GHMM}(w_k(l)|\boldsymbol{w}) = \mathcal{B}(k) \\ \frac{J_k^*(l)}{F} & : \ \mathrm{GHMM}(w_k(l)|\boldsymbol{w}) \neq \mathcal{B}(k) \end{cases} \quad (5)$$

The use of this context-sensitive confidence value allows the replacement of the uniform change of word-hypotheses with the described *weighted* change by calculating the probability $P(w_k(l)) = \bar{J}_k(l)/\sum_{l_0=1}^{L} \bar{J}_k(l_0)$.

5. EXPERIMENTAL RESULTS

In this section we show the learning ability of our approach by evaluating the performance of four different system variants:

System variant 1 does not contain any restrictions and initializes the primitive nodes with randomly chosen word-hypotheses.

System variant 2 does not contain any restrictions but uses a *heuristic* initialization for the assignment of word-hypotheses to primitive nodes.

System variant 3 contains the manually entered restrictions and uses the heuristic initialization.

System variant 4 is based exclusively on the confidence values introduced in section 4, both for initialization and weighted changes of word-hypotheses. This system variant was trained by removing the existent, manually modelled restrictions out of system variant 3 and learning them from scratch by means of the set of classified utterances.

For our experiments we use 6712 spontaneous utterances collected using a version of the EVAR system connected to the public telephone network. There are transliterations and a pragmatic labeling available for each utterance. Training is done with 6385 utterances and for testing we use the remaining 327 ones.
The quality of an interpretation is measured according to the amount of correct pragmatic information units found. In Table 1 the percentage of the correctly analyzed pragmatic unit *place of arrival* is shown for all four system variants. The results were achieved after n iterations on five processors. It can be seen that system variant 4 which utilizes the new confidence values is superior to all the other systems. This proofs that our approach is able to learn the linguistic restrictions and that the confidence values are doing very well for the initialization and the weighted change of word-hypotheses.
The results for the other important pragmatic units *place of departure* and *date/time of departure* (not shown in Table 1) are less convincing. The results after performing 25 iterations on five processors for the pragmatic unit

	system variant			
n	1	2	3	4
1	34.2	79.7	92.5	94.6
10	41.0	79.7	93.9	94.9
25	48.8	76.6	94.6	95.6

Table 1. Percentage of correctly analyzed pragmatic unit *place of arrival*. n denotes the number of iterations. The four system variants are described in the text.

- *place of departure* are 95.8% for system variant 3 and 93.3% for system variant 4.
- *date/time of departure* are 71.3% for system variant 3 and only 41.7% for system variant 4.

The overall result (containing all pragmatic units) is 90.8% for system variant 3 and 85.1% for variant 4. The reason for the lower results of system variant 4 is due to some technical aspects: Most of the functions referenced by the attributes, the relations and the confidence measure of the concepts are still based on the manually entered restrictions and the binary decision measure (cf. Section 4). These functions still have to be adapted to the new rating function.

6. CONCLUSION AND FUTURE WORK

In this paper we proposed the use of a statistically based method for automatically learning domain specific knowledge within semantic networks. The experimental results show that our approach is well suited for this task and that it can significantly speed up the adaption of a semantic network knowledge base to new domains by automatically learning a lot of the necessary domain-specific knowledge.
In our future work we will concentrate on the necessary modifications to take full advantage of our approach's potential. We further want to adapt our system to a new domain with extensive use of our proposed approach.

REFERENCES

[1] Geert Jan Wilms. Using a Hybrid System of Corpus– and Knowledge–Based Techniques to Automate the Induction of a Lexical Sublanguage Grammar. In *Proc. of the Int. Conf. on Computational Linguistics*, volume 2, pages 1163–1166, Kopenhagen, August 1996.

[2] Rens Bod. *Beyond Grammar — An Experience–Based Theory of Language*, volume 88 of *CSLI Lecture Notes*. CSLI Publications, California, 1998.

[3] Fischer, J. and Niemann, H. and Noeth, E. A Real–Time and Any–Time Approach for a Dialog System. In *Proc. International Workshop Speech and Computer (SPECOM'98)*, pages 85–90, St.–Petersburg, 1998.

[4] H. Niemann, G. Sagerer, S. Schröder, and F. Kummert. ERNEST: A semantic network system for pattern understanding. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 9:883–905, 1990.

[5] Fischer, J. and Haas, J. and Nöth, E. and Niemann, H. and Deinzer, F. Empowering Knowledge Based Speech Understanding through Statistics. In *ICSLP*, volume 5, pages 2231–2235, Sydney, Australia, dez 1998.

[6] G. Box and G. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, 1973.

[7] H. Steinhaus. The problem of estimation. *Annals Math. Statistic*, 28:633–648, 1957.

[8] J. Haas, J. Hornegger, E. Nöth, and H. Niemann. A Probabilistic Approach for the Semantic Analysis. In *Proc. of the AIII Workshop on Artificial Intelligence in Industry*, pages 422–430, 1998.