BMBF

Verbmobil
Verbundvorhaben

Some Recognition Results for INTARC 2.0

Hans Weber, Jörg Spilker,
Günther Görz

Universität Erlangen-Nürnberg

Vm
Report 187
Dezember 1996

Dezember 1996

Hans Weber, Jörg Spilker, Günther Görz

IMMD VIII – Künstliche Intelligenz
Universität Erlangen-Nürnberg
Am Weichselgarten 9
D-91058 Erlangen

Tel.: (+49 9131) 85 - 9907 - 118
Fax: (+49 9131) 85 - 9907 - 05

Gehört zum Antragsabschnitt: 15.7: Architektur integrierter Parser für gesprochene Sprache

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 H9 und 01 IV 101 A/O gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Some Recognition Results for INTARC 2.0

Hans Weber, Jörg Spilker, Günther Görz

January 23, 1997

1 General Remarks

This report presents some word recognition results on the speech–to–speech translation system INTARC 2.0. Our research goal was to build a system with a cognitive oriented architecture. The main topics are incremental, time-synchronous and interactive processing. All modules work on the same time segment processing the signal from left to right. Analyses – even partial ones – are passed as soon as possible. Figure 1 shows the overall structure of the system. Details of the architecture of parts of INTARC 2.0 can be found in [1][1].

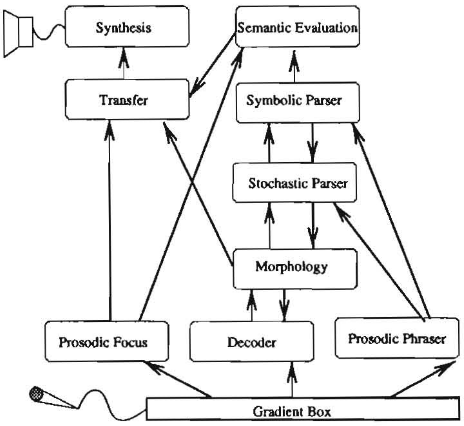

Figure 1: The flow of data in INTARC 2.0

For test data we used 20 utterances from the VERBMOBIL Collection, which is spontaneous dialogue speech.

[1] The cited paper concentrates on the left right incremental parser and the probabilistic grammar model.

2 Conditions

2.1 The Modules Involved

Not all of the modules and configurations of INTARC 2.0 have been tested yet. This paper concentrates on five modules.

- Word–Part Recognition Module, University of Hamburg
- Morphological Analyzer (Morphy), University of Bielefeld
- Probabilistic Lattice Parser, University of Erlangen
- Symbolic Semantic Parser, DFKI Saarbrücken
- Prosodic Phrase Boundary Module, University of Bonn

We tested three different module configurations

DM Decoder, Morphy (acoustic word recognition)

DMP Decoder, Morphy, Lattice Parser (word recognition in parsed utterances)

DMPS Decoder, Morphy, Lattice Parser, Semantic Module (word recognition in understood utterances)

The configurations correspond to successively harder tasks, namely recognize, analyze and understand.

2.2 Data Flow

The INTARC-System has two modes of operation resulting in different data flow. These two are

TD Bottom up and top down data flow between parser to decoder

BU The standard bottom up data flow from decoder to parser only

In top down mode the parser uses a precompiled prediction table for possible continuations of the actual analyses. The prediction table is built up with the original DFKI-Grammar and the stochastic grammar model. For every rule and daughter we determine the words possible in the DFKI-Grammar and look up the corresponding probability. A fixed beam is used to prune all words with low probabilities.

2.3 The models used

2.3.1 Unification Grammar

The unification grammar used in the experiments consists of 700 lexical entries and 60 rules. It had originally been written in the Type Definition Language of DFKI Saarbrücken and was compiled into the ASL-Features Formalism suitable for the training procedure of the grammar derivation model (GM).

In the INTARC 2.0 lattice parser a context–free approximation of the Saarbrücken Grammar (SG) is processed. The approximation grammar corresponds to a second order Markov model.

2.3.2 Grammar Model (GM)

We used a variant of Inside-Outside training to estimate a model of the unification grammar derivations. It is a trigram model similar to PCFG but with more context with respect to predecessor rules in a derivation.

2.3.3 Bigram Model

In all of the experiments a word form based bigram model has been used, trained by Kai Hübener at University of Hamburg[2]. The model perplexity is 100.

2.3.4 Prosody Trigram

We used a trigram model of word categories and phrase boundary categories to score combinations of words with phrase boundary hypotheses supplied by the prosody module. The model we took has been developed by Michael Lehning at the University of Braunschweig.

2.4 The data

The most prominent problem was to find enough data covered by the grammar both for training the unification grammar model as well as for testing the whole system.

Since the Saarbrücken Grammar was written in a linguistic style, we had to add two extra recursive rules allowing for multi sentence analyses. As has been noted by others before, those rules lead to a great loss in performance with respect to time and space resources. Nevertheless, we decided to neglect performance in order to achieve an acceptable coverage without using a lot of corpus specific rules.

We chose the VM Dialogues n001k, n002k, n009k and n011k to train the grammar model.

Many of the turns were split into smaller chunks manually, resulting in a training corpus where no turn is longer than one sentence. 70 percent of the resulting corpus was parsable by the Saarbrücken Grammar and was used to train the grammar probabilities.

[2] VERBMOBIL partner 15.2.

As test data we used the following 20 utterances which were designed manually. The design procedure was: We chose some dialogue sentences from the VM–corpus. In these turns we replace all out–of–vocabulary words with similar meaning in–vocabulary–words. Finally the turns were read in a non–spontaneous style.

```
<SIL> GUTEN TAG HERR KLEIN
<SIL> K-ONNEN WIR UNS AM MONTAG TREFFEN
<SIL> JA DER MONTAG PA-ST MIR NICHT SO GUT
<SIL> JA DANN TREFFEN WIR UNS DOCH AM DIENSTAG
<SIL> AM DIENSTAG HABE ICH LEIDER EINE VORLESUNG
<SIL> BESSER W-ARE ES BEI MIR AM MITTWOCH MITTAGS
<SIL> ALSO AM MITTWOCH UM ZEHN BIS VIERZEHN UHR HABE ICH ZEIT
<SIL> DANN LIEBER GLEICH NACH MEINEM DOKTORANDENTREFFEN
<SIL> WOLLEN WIR UNS NICHT LIEBER IN MEINEM B-URO TREFFEN
<SIL> NA JA DAS W-URDE GEHEN
<SIL> JA HERR KLEIN WOLLEN WIR NOCH EINEN TERMIN AUSMACHEN
<SIL> VIELLEICHT GINGE ES AM  MITTWOCH IN MEINEM B-URO
<SIL> DAS IST DER VIERZEHNTE MAI
<SIL> AM MITTWOCH DEN VIERZEHNTEN PA-ST ES MIR NICHT SO GUT
<SIL> AM DIENSTAG IN DIESER WOCHE H-ATTE ICH NOCH EINEN TERMIN
<SIL> ALSO DANN AM DIENSTAG DEN DREIZEHNTEN MAI
<SIL> VORMITTAGS ODER AM NACHMITTAG
<SIL> JA MACHEN SIE DOCH EINEN VORSCHLAG
<SIL> JA DANN LASSEN SIE UNS DOCH DEN VORMITTAG NEHMEN
<SIL> JA GUT TSCH-U-S
```

2.5 Measuring Performance

We used the NIST scoring program for word accuracy to gain comparable results. By doing this we gave preference to a well known and practical measure although we know that it is in some way inadequate.

In a system like INTARC 2.0, the analysis tree is of much more importance than the recovered string. In VERBMOBIL, the global research goal is translation of spontaneous speech, so a good semantics for a string with word errors is more important than a good string with a completely wrong reading.

Second, the grammar scores have only indirect influence on the string. Their main function is picking the right tree. For the Saarbrücken Grammar there exists no tree bank with correct readings for our test data. So we had no opportunity to measure something like a “tree recognition rate” or “rule accuracy”[3].

The word accuracy results in DMP and DMPS can not be compared to word accuracy as usually applied to an acoustic decoder in isolation. The DM values can be compared in this way.

In DMP and DMPS we counted only those words as recognized which could be built into a valid parse from the beginning of the utterance. Words to the right, which could not be integrated into a parse, were counted as deletions —

[3] Note that the word string is contained in the tree but only as a part of it.

although they might have been correct in standard word accuracy terms. Our evaluation method is much harder than standard word accuracy, but it appears to be a good approximation to "rule accuracy". We think that what cannot be parsed is not usable in a VERBMOBIL system, so we have to count it as an error.[4]

The difference between DMP and DMPS is, that a tree produced by the statistical approximation grammar can be ruled out when rebuilt by with unification operations in semantic processing. The loss in recognition performance from DMP and DMPS corresponds to the quality of the statistical approximation. If the approximation grammar had a 100 percent tree recognition, there would be no gap between DMP and DMPS.

3 Test results

The recognition rates of the three configurations are measured in three different contexts. The first row of the following table shows the rates of normal bottom up processing. In the second row the results of the phrase boundary detector are used for disambiguation for syntax and semantics. The third row shows the results of the system in top down mode. Here no semantic evaluation is done because the top down predictions only affect the interface between syntax module and decoder.

	DM	DMP	DMPS
Word Accuracy	93,9%	83,3%	47,5%
WA with phrase boundary	93,9%	84,0%	48,6%
WA in TD-Mode	94.0%	83,4%	–

4 Conclusions

Splitting composite nouns to reduce the decoder's lexicon shows good results. The searching and rebuilding performed by the morphology module is implemented as a finite state automaton, so there is no great loss in performance. Incremental decoding is as good as as the standard decoding algorithms, but the lattices are up to ten times larger. This causes a performance problem for the parser. Our approach is to use an approximation of a HPSG-Grammar for searching. So the syntactic analysis becomes more or less a second decoding step. By regarding a wider context, we even reduce the recognition gap between syntax and semantics in comparison with our previous unification-based syntax parser (see [2, 3]). With respect to a really usable system the tree-recognition rate must be improved. This can be done by a bigger training set. The actually used dialogs contains only 83 utterances. A second improvement can be

[4] On the long run our results should be compared with those of a standard serial architecture. This means to take the output of an n-best-decoder and delete everything which is not parsable from that output. Word accuracy should be measured for the remaining strings. Alternatively a similar procedure could be applied to standard lattice parsing.

achieved by a larger context during training to get a better approximation of the trees built by the HPSG-grammar.

Prediction of words seems to have no influence on the recognition rate. This is a consequence of the underlying domain. The HSPG-Grammar is written for spontaneous speech, so nearly every utterance should be accepted. The grammar gives no restrictions on possible completions of an utterance. Restrictions can be only obtained by a narrow beam-bound when compiling the prediction table. But this leads to a lower recognition rate because some correct words are pruned.

References

[1] Walter Kasper, Hans-Ulrich Krieger, Jörg Spilker, and Hans Weber. From word hypotheses to logical form: An efficient interleaved approach. In D. Gibbon, editor, *Natural Language Processing and Speech Technology. Results of the 3rd KONVENS Conference*, pages 77–88. Mouton de Gruyter, Berlin, 1996.

[2] H. Weber. *LR-inkrementelles probabilistisches Chartparsing von Worthypothesenmengen mit Unifikationsgrammatiken: Eine enge Kopplung von Suche und Analyse.* PhD thesis, Universität Hamburg, FB Informatik, Dezember 1994. Auch: VERBMOBIL Report 52, Universität Erlangen-Nürnberg, IMMD 8, 1995.

[3] Hans H. Weber. Time-synchronous chart parsing of speech integrating unification grammars with statistics. *Proceedings of Twente Workshop on Speech and Language Engineering*, pages 107–120, December 1994.