Statistical Approaches to Language" Workshop at the 32nd Annual Meeting of the ACL, 1994

T. Kemp: *Data-Driven Codebook Adaptation in phonetically tied SCHMMS.* to appear in Proc. ICASSP 95

T. Sloboda: *Dictionary Learning: Performance through Consistency.* to appear in Proc. ICASSP 95

P. Geutner: *Using Morphology towards better Large Vocabulary Speech Recognition Systems.* to appear in Proc. ICASSP 95

U. Bodenhausen: *Automatic Structuring of Neural Networks for Spatio-Temporal Real-World Applications.* Ph.D thesis, University of Karlsruhe, June 1994

J.-L. Gauvin, L.-F. Lamel, G. Adda and M. Adda-Decker: *The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task.* Proc. ICASSP 94, vol. 1, pp. 557-560

W. Ward: *Understanding Spontaneous Speech: The Phoenix System.* Proc. ICASSP 91, vol. 1, pp. 365–367

G. Gazdar, E. Klein, G.K. Pullum, and I.A. Sag: *Generalized Phrase Structure Grammar.* Blackwell Publishing, Oxford, England and Harvard University Press, Cambridge, MA, USA, 1985

R. Kaplan and J. Bresnan: *Lexical-functional grammar: A formal system for grammatical representation.* In *The Mental Representation of Grammatical Relations*, pp. 173-281. The MIT Press, Cambridge, MA, 1982.

C. Pollard and I. Sag: *An Information-Based Syntax and Semantics.* CSLI Lecture Notes No.13, 1987.

C. Nakatani and J. Hirschberg: *A Speech–First Model for Repair Identification in Spoken Language Systems.* Proc. of the ARPA Workshop on Human Language Technology, March 1993

A.-E. McNair and A. Waibel: *Improving Recognizer Acceptance through Robust, Natural Speech Repair.* Proc. ICSLP 94, vol. 3, pp. 1299-1303

S.-R. Young and W. Ward: *Learning New Words from Spontaneous Speech.* Proc. ICASSP 93, vol. 2, pp. 590-591

Robust Speech Repair

After locating and highlighting erroneous sections in the recognizer hypothesis, misrecognitions are corrected.

The *spoken hypothesis correction* method uses N-Best lists for both the initial utterance and the respoken section. The N-Best list for the highlighted section of the initial utterance is rescored using scores from decoding the secondary utterance. Depending on the quality of the N-Best lists, most misrecognitions can be corrected.

The *spelling hypothesis correction* method requires the user to spell the highlighted erroneous section. A spelling recognizer decodes the spelled sequence of letters. By means of a language model we restrict the sequence of letters to alternatives found among the N-Best from the located section.

To date, we have evaluated our methods over sentences from the Resource Management task. Table 7 shows the improvements in sentence accuracy, based on recordings from one speaker of the February and October 1989 test data. We selected a subset of erroneous utterances; therefore the accuracy of the baseline system is significantly lower than the 94% performance our system achieves on the whole test set. The results indicate that repeating or spelling a misrecognized subsection of an utterance can be an effective way to repair recognition utterances.

No Repair (baseline)	63.1%
Respeak	83.8%
Spell	88.5%
Respeak + Spell	89.9%

Table 7: Improvement of Sentence Accuracy by Repair

Conclusions

We described JANUS-2, our multilingual spoken language translation system. We introduced different learning approaches that reduce hand-tuning efforts, yield better word accuracy and even accelerate recognition speed. All of these techniques can be applied in several languages and help making significant advances towards building a multilingual translation system for spontaneous human-to-human dialogs. Beyond recognition of spontaneous speech JANUS-2 provides a framework for investigating important areas like robust parsing, machine translation of spoken language and developing methods to recover from recognition and parsing errors.

Acknowledgements

The German speech recognition engine was funded by grant 413-4001-01IV101S3 from the German Federal Ministry of Education, Science, Research and Technology (BMBF) as a part of the VERBMOBIL project. The English and Spanish speech translation components were funded in part by grants from the Advanced Research Project Agency, the US Government, Siemens Corporation and ATR Interpreting Telecommunications Research Labs of Japan. We gratefully acknowledge their support. The views and conclusions contained in this document are those of the authors.

References

L. Osterholtz, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, A. Waibel and M. Woszczyna: *Testing Generality in JANUS: A Multi-Lingual Speech to Speech Translation System.* Proc. ICASSP 92, vol. 1, pp. 209-212

M. Woszczyna, N. Coccaro, A. Eisele, A. Lavie, A. McNair, T. Polzin, I. Rogina, C.P. Rose, T. Sloboda, M. Tomita, J. Tsutsumi, N. Aoki–Waibel, A. Waibel and W. Ward: *Recent Advances in JANUS: A Speech Translation System.* Proc. EUROSPEECH 93, vol. 2, pp. 1295-1298

B. Suhm, P. Geutner, T. Kemp, A. Lavie, L. Mayfield, A.E. McNair, I. Rogina, T. Sloboda, W. Ward, M. Woszczyna and A. Waibel: *JANUS: Towards Multilingual Spoken Language Translation.* DARPA Speech and Natural Language Workshop 1994

W. Wahlster: *Verbmobil: Translation of Face-To-Face Dialogs.* DFKI, November 1993

C-STAR - Consortium for Speech Translation Research: *Organization and Goals.* Unpublished Notes, München, June 1994

H. Hild and A. Waibel: *Speaker-Independent Connected Letter Recognition With a Multi-State Time Delay Neural Network.* Proc. EUROSPEECH 93, vol. 2, pp. 1481-1484

O. Schmidbauer and J. Tebelskis: *An LVQ based Reference Model for Speaker–Adaptive Speech Recognition.* Proc. ICASSP 92, vol. 1, pp. 441-445

I. Rogina and A. Waibel: *Learning State–Dependent Stream Weights for Multi–Codebook HMM Speech Recognition Systems.* Proc. ICASSP 94

B. Suhm and A. Waibel: *Towards Better Language Models for Spontaneous Speech.* Proc. ICSLP 94, vol. 2, pp. 831-834

A. Lavie and M. Tomita: *GLR* - An Efficient Noise-skipping Parsing Algorithm for Context-free Grammars.* Proceedings of Third International Workshop on Parsing Technologies, 1993, pp. 123–134

B. Suhm, L. Levin, N. Coccaro, J. Carbonell, K. Horiguchi, R. Isotani, A. Lavie, L. Mayfield, C. Pennstein-Rosé, C. Van Ess-Dykema and A. Waibel: *Speech–Language Integration in a Multi–Lingual Speech Translation System.* Workshop on Integration of Natual Language and Speech Processing, AAAI-94, Seattle

F.-D. Buø, T.-S. Polzin and A. Waibel: *Learning Complex Output Representations in Connectionist Parsing of Spontaneous Speech.* Proc. ICASSP 94, vol. 1, pp. 365-368

L. Mayfield, M. Gavalda, W. Ward and A. Waibel: *Concept-Based Speech Translation.* to appear in Proc. ICASSP 95

C. Penstein Rosé and A. Waibel: *Recovering From Parser Failures: A Hybrid Statistical/Symbolic Approach.* to appear in "The Balancing Act: Combining Symbolic and

2. *The Linguistic Feature Labeler* attaches features and feature values (if applicable) to these chunks. There is a classifier for each feature, which finds zero or one atomic value. Since there are many features, each chunk may get none, one or several pairs of feature and atomic values. As a feature normally only occurs at a certain chunk level, the classifier is specialized to decide on a particular feature at a particular chunk level. This specialization prevents the learning task from being too complex, thus keeping it easy to learn.

3. *The Chunk Path Finder* determines how a chunk relates to its parent chunk. It has one classifier per chunk level and chunk path element.

The following English sentence will illustrate the work of the parser:

Can you meet in the morning

The Chunker segments the sentence before passing it to the Linguistic Feature Labeler, which adds semantic labels (shown in **boldface** below):

```
(((speech-act *suggest)
  (sentence-type *query-if))
  ((frame *free))
   ((                              can))
   (((frame *you))
    (                              you))
   ((                              meet))
   ((frame *special-time))
    (                              in)
    (((specifier definite))        the)
    (((time-of-day morning))       morning)))
```

The Chunk Path Finder then adds paths, where appropriate (shown in **boldface**):

```
([](( speech-act *suggest)
    ( sentence-type *query-if))
    ([]( frame *free))
      ([]([]                              can))
      ([who](( frame *you))
        ([]                               you))
      ([]([]                              meet))
      ([when](( frame *special-time))
        ([]                               in)
        ([](( specifier definite))        the)
        ([](( time-of-day morning))       morning)))
```

Converting this into feature structure, we get the following semantic feature structure (ILT):

```
(((speech-act *suggest)
  (sentence-type *query-if)
  (frame *free)
  (who ((frame *you)))
  (when ((frame *special-time)
         (specifier definite)
         (time-of-day morning)))))
```

Based on this ILT representation, utterances in the target language can be generated.

Handling Unreliability

Since a speech translation system involves interaction between two human users, the system should provide methods for adaptive recovery from misrecognitions, miscommunication and mistranslations. First results in this direction are described here.

We have developed a speech interface for repairing recognition errors by simply respeaking or spelling a misrecognized section of an utterance. While much speech "repair" work has focused on repairs within a single spoken utterance (Nakatani & Hirschberg 1993), we are concerned with the interactive repair of errorful recognizer hypotheses (McNair & Waibel 1994).

Identifying Errors

To be able to repair an error, its location has to be determined first. We pursue two strategies to identify misrecognitions as subpieces of the initial recognizer hypothesis.

The *automatic subpiece location* technique requires the user to respeak only the errorful subsection of the (primary) utterance. This (secondary) utterance is decoded using a vocabulary and language model limited to substrings of the initial erroneous hypothesis. Thus, the decoding identifies the respoken section in the hypothesis. Preliminary testing showed that the method works poorly if the subpiece to be located is only one or two words long. However, this drawback is not severe since humans tend to respeak a few words around the error.

A second technique uses *confidence measures* to determine for each word in the recognizer's hypothesis whether it was misrecognized. First, we applied a technique similar to Ward (Young & Ward 1993), which turns the score for each word obtained during decoding into a confidence measure by normalizing the score and using a Bayesian updating technique based on histograms of the normalized score for correct and misrecognized words. Since we found this not to work well on our English scheduling task, we are currently developing different methods to compute confidence measures based on decoder, language model and parse scores.

Concept Based Speech Translation

The basic premise of the concept based approach is that the structure of the information conveyed is largely independent of the language used to encode it. Our system tries to model the information structures inherent in a task, e.g. the scheduling task, and the way these structures are represented through words in various languages. This system is an extension of the Phoenix Spoken Language System (Ward 1991). It uses the Phoenix parser to parse input into slots of semantic frames, and then uses these frames to generate output in the target language.

The Parser Unlike individual words, semantic units used in a task domain are not language specific. Based on transcriptions of scheduling dialogs, we have developed a set of fundamental semantic units in our parse which represent the different concepts a speaker would use. For instance, a typical *temporal* token could have *date* as subtoken, which could in turn consist of *month* and *day* subtokens. The *temporal* token could be part of a statement of unavailability.

In contrast to previous speech translation systems, we presently don't perform syntactic analysis. Speaker utterances, as decoded by the recognizer, are parsed into semantic chunks which are concatenated without grammatical rules. This approach is particularly well suited to parsing spontaneous speech, which is often ungrammatical and subject to recognition errors. This approach is more robust than requiring well-formed input and the reliance on syntactic cues provided by short function words such as articles and prepositions.

The Generator The generation component of the system is a simple left-to-right processing of the parsed text. The translation grammar consists of a set of target language phrasings for each token, including lookup tables for variables like numbers and days of the week. When a lowest-level token is reached in tracing through the parse, a target language representation is created by replacing tokens with templates for the parent token, according to the translation grammar. The result is a meaningful, although terse translation, which emphasizes communicating the main point of an utterance. An example is illustrated in figure 3.

Results We have implemented this system for bidirectional translation between English, German and Spanish in our scheduling task. Table 6 shows the performance of parser and subsequent generator on transcribed data. Evaluation of the system based on speech decoded by the JANUS-2 recognizer is still underway.

One disadvantage of this approach is the telegraphic and repetitive nature of the translations. This could be overcome by providing multiple translation options for individual tokens in the target language module, different levels of politeness, etc. However, at present we feel that it is sufficient for intelligible communication.

Original utterance:

THAT SATURDAY I'M NOT SURE ABOUT BUT YOU SAID YOU MAY BE BACK IF YOU THINK YOU'LL BE BACK THIS SUNDAY THE TWENTY EIGHTH I COULD SEE YOU AFTER ELEVEN AM ON THAT IF YOU'RE BACK

Translated:

Saturday that's not so good for me Sunday the twenty eighth works for me after eleven a.m. (ENGLISH)

El sábado no me va demasiado bien pero el domingo veintiocho me va bien después de las once de la mañana. (SPANISH)

Samstag könnte ich nur zur Not aber Sonntag der achtundzwanzigste geht bei mir ganz gut nach elf Uhr morgens. (GERMAN)

Figure 3: Translation Example

	Parsed from		Translated into
	token	utterance	utterance
English	95.6%	90.0%	90.2%
German	92.4%	89.6%	87.3%
Spanish	88.8%	58.3%	82.2%

Table 6: End–to–End evaluation on transcribed data

Connectionist Parser

Alternatively to a slot based parser, also a connectionist parser can be used. The major drawback of connectionist parsers, compared with symbolic parsers, has been the lack of detailed linguistic information in the output. We present a connectionist natural language parser that learns to produce feature structures (Gazdar et al. 1985; Kaplan & Bresnan 1982; Pollard & Sag 1987), the most frequently used representation scheme in computational linguistics. The most outstanding advantage of this parser is its learnability of complex feature structures, by automatically splitting the overall problem into several small classification tasks which are learnable (Divide and conquer). Each of these smaller tasks is then learned by a separate backpropagation network, using constructive learning. Constructive learning ensures an optimal network architecture, being as small as possible and achieving the highest possible performance.

Parser principle The parser consists of three main parts:

1. *The Chunker*, composed of three networks, splits an input sentence into chunks. The first network finds regular expressions, such as numbers. Numbers are classified as being ordinal or cardinal numbers. They are presented as words to the following networks. The next network arranges words to phrases. The third network puts together phrases to clauses. In total, we get four levels of chunks: words, phrases, clauses and sentences.

matrix of the speech recognizer used. First, phonetic transcriptions for all appearances of each word are generated by the help of a phoneme recognizer. Then, variants which are infrequent or which would lead to erroneous training of confusable phonemes are eliminated. Finally, the acoustic models are retrained allowing for the newly acquired pronunciations variants.

As can be seen in table 4, our algorithm for adapting and adding phonetic transcriptions to a dictionary improves the recognition accuracy of the decoder significantly and, for a context independent recognition system, leads to performance that is comparable to the context dependent results (cf. table 3). The baseline decoder for these experiments uses 69 context independent phoneme models. Evaluation using context dependent models is in progress.

Dictionary	Word Accuracy
baseline	61.7%
adapted	65.6%

Table 4: Results for Dictionary Learning (GSST)

Morpheme–Based Language Models

Comparing various languages like English, Spanish and German, it can be easily seen that German and Spanish differ from English by an outstanding number of inflections and compound words. Due to this fact dictionaries for morphologically rich languages grow much faster with increasing database size, compared to English (cf. figure 2). One way to limit this growth with an increasing amount of training data is to use smaller base units than words within the recognition process.

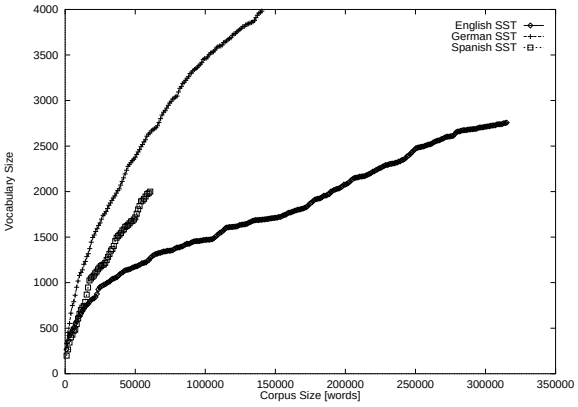

Figure 2: Vocabulary Growth

Concerning German, three different ways of word decomposition have been evaluated:

1. strictly morpheme-based decomposition, e.g.:
   - weggehen → weg-geh-en[1] (*to go away*)
   - Spracherkennung → Sprach-er-kenn-ung (*speech recognition*)
2. decomposition in root forms:
   - weggehen → weggeh@ (*to go away*)
   - Dialoge → Dialog@ (*dialogs*)
3. combination of strictly morpheme-based decomposition and root forms:
   - weggehen → weg-geh@ (*to go away*)

Table 5 shows dictionary size and recognition accuracy using the respective decomposition methods, based on 250 GSST dialogs. As can be seen, all decomposition methods reduce vocabulary size. The impact on recognition accuracy is small, but the morpheme-based approach outperforms the open-vocabulary baseline system. The only small improvement may be due to the fact that the acoustic confusability increases when using smaller recognition units and thus deteriorates the gain in the language model. In a real interface, however, this reduction in vocabulary growth leads to a reduction of new words, thus reducing the word error rate, and smaller dictionaries also accelerate recognition speed significantly. Further research will focus on finding more efficient and acoustically less confusable decompositions automatically, and also test the impact on translation.

	Dictionary	Accuracy
Baseline (closed-vocabulary)	3085	66.9%
Baseline (open-vocabulary)	3062	64.7%
Morphemes	2204	65.8%
Root Forms	3062	63.5%
Combined	2998	65.1%

Table 5: Comparison of Decomposition Methods (GSST)

Speech Translation

We are developing various translation schemes like a generalized robust LR parser (Lavie & Tomita 1993), statistical grammar inference, a concept based translation approach (Mayfield et al. 1995) and a connectionist parsing approach (Buø, Polzin, & Waibel 1994). In this paper the two latter will be described.

[1] Hyphens are used for clarification purposes as decomposition markers only and do not appear in the actual German spelling.

approach leads to improved performance with appropriate weighting of the output from each strategy.

Recognition Performance Analysis

The baseline JANUS-2 recognizer can be described as follows:

- *Preprocessing*: LDA on melscale fourier spectrum and additional acoustic features (power, silence)
- *Acoustic modeling*: LVQ-2 or phonetically tied SCHMM, explicit noise models
- *Decoder*: Viterbi search as first pass, followed by a word-dependent N-Best search, standard word bigram language model, word lattice output

Current recognition results on the English, German and Spanish Spontaneous Scheduling Task (ESST, GSST, SSST) can be seen in table 2.

	ESST	GSST	SSST
Word Accuracy	66%	69.9%	61%

Table 2: JANUS-2 recognition performance

The low absolute recognition accuracies are due to the challenging nature of human-to-human spontaneous speech. Recent evaluations on the Switchboard task confirm that human-to-human dialogs are much more difficult to recognize than human-machine spontaneous speech (like ATIS). Current state-of-the-art systems achieve word accuracies between 30% and 50% on the Switchboard database.

Perplexities range between 35 and 90 for ESST, SSST and GSST, and somewhat over 100 for Switchboard. Additionally, human-to-human dialogs are significantly more disfluent (Suhm et al. 1994a). Large variations in speaking rates and strong coarticulation between words contribute considerably to the difficulty of recognizing human-to-human spontaneous speech.

Different Learning Approaches

The following three sections will describe efforts and results of improving the recognition component along its major knowledge sources: acoustic models (Kemp 1995), dictionary (Sloboda 1995) and language models (Geutner 1995).

Data–Driven Codebook Adaptation

The performance of a parametric classifier is always dependent on the adequacy of the underlying model assumptions. In speech recognition with HMM-based systems, usually the model assumption for the distribution of the data in feature space is the sum of N multivariate gaussian distributions. Whereas this model assumption can be shown to cover all possible distributions, this holds only if the number of gaussians is chosen correctly. Mainly governed by practical concerns, in most speech recognition systems this number is often chosen to be the same power of 2 for each of the different phonemes that have to be modelled, meaning that a fixed number of codebook vectors is assigned to each of the phonemes. However, as the available training data differs between phonemes, and the size of the feature space covered by the different phonemes varies greatly, constant codebook size leads to suboptimal allocation of resources.

We therefore suggest methods aimed at automatic optimization of the number of parameters for the semi-continuous phonetically tied HMM used in JANUS-2. We have developed (Kemp 1995) two different algorithms to adapt the codebook size of each phoneme according to the amount and the distribution of the training data similar to (Bodenhausen 1994). Basically, both algorithms start with one gaussian and during training the amount of parameters is incremented until some quality criterion determines when to stop the process of increasing the codebook. We compared a *variance* criterion based on the average distance between data points and their nearest codebook vector with a *prediction* criterion which tries to capture how well the modeling of the recognizer can predict unseen data.

Model	Codebook Size	Word Accuracy
baseline	4600	66.9%
variance	4201	69.9%
prediction	1677	67.8%

Table 3: Results for Codebook Adaptation (GSST)

Table 3 compares recognition accuracies and codebook sizes of the baseline models with models automatically adapted using the variance and prediction criterion. As can be seen, the more efficient parameter allocation when adapting codebook sizes leads to significant error reduction if the same number of parameters is used. Furthermore, the number of parameters can even be reduced by 60% with still better performance than the baseline system.

Dictionary Learning

Due to the enormous variability in spontaneous human-to-human dialogs, creating adequate dictionaries with alternative pronunciations is crucial (Gauvin et al. 1994). However, hand tuning and modifying dictionaries is time consuming and labor intensive. Pronunciations of a word should be chosen according to their frequency and also modifications of the dictionary should not lead to higher phonetic confusability after retraining. Therefore we have proposed (Sloboda 1995) a data-driven approach to improve existing dictionaries and automatically add new words and pronunciation variants whenever needed.

The learning algorithm requires transcriptions for the whole training set and a phoneme confusability

nally, we report on efforts to detect erroneous system output and provide interactive methods to recover from such errors.

JANUS Overview

Data Collection

Data collection to establish a large database of spontaneous human-to-human negotiation dialogs in English and German has started about 18 months ago. In the meantime, several sites in Europe, the US and Asia have adopted the scheduling task under several research projects and funding sources. Since the same calendars and data collection protocols are used the data elicited shares the same domain and procedural constraints.

English Scheduling		
	dialogs	words
recorded	1984	505 K
transcribed	1826	460 K
German Scheduling		
	dialogs	words
recorded	734	158 K
transcribed	534	115 K
Spanish Scheduling		
	dialogs	words
recorded	340	79 K
transcribed	256	70 K
ATIS3		
transcribed	n/a	250 K

Table 1: Comparison of Databases

Table 1 summarizes the current status of data collection. Since scheduling utterances typically consist of more than one sentence, there is already more data available for English scheduling than ATIS [1]. More data collection will establish databases in size at least comparable to ATIS for all languages.

System Overview

The main system modules are speech recognition, parsing, discourse processing, and generation of target language output. Each module is language-independent in the sense that it consists of a general processor that applies independently specified knowledge about different languages.

The recognition module decodes the speech in the source language into a list of sentence candidates, represented either as a word lattice or N-Best list. At the core of the machine translation components is a language independent representation of meaning (ILT), which is extracted from the recognizer output by the parsing module. As last step, this language independent representation is sent to the generator to be translated into any of the target languages. Figure 1 shows the system architecture.

After parsing, a discourse processor can be used to put the current utterance in the context of previous utterances. Based on the current discourse state, speech and natural language processing system components can be integrated to resolve parsing ambiguities and dynamically adapt the vocabulary and language model of the recognizer.

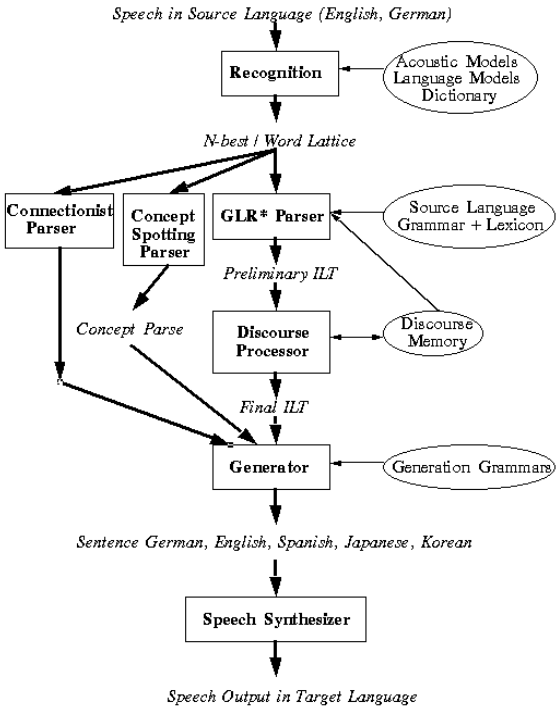

Figure 1: System Architecture

We explore several approaches for the main processes. We are experimenting with TDNN's, MS-TDNN's (Hild & Waibel 1993), MLP's, LVQ (Schmidbauer & Tebelskis 1992) and HMM's (Rogina & Waibel 1994; Kemp 1995) for acoustic modeling. We are using n-grams, word clustering, automatic phrase detection (Suhm & Waibel 1994) and morpheme-based approaches for language modeling (Geutner 1995). Statistically trained skip parsing (Lavie & Tomita 1993; Suhm et al. 1994a), neural net parsing (Buø, Polzin, & Waibel 1994) and concept spotting parsing (Mayfield et al. 1995) are being applied for extracting the meaning. Also statistical models as well as plan inferencing for identification of the discourse state (Rosé & Waibel 1994) are being used. This multi-strategy

[1] The approximately 18000 utterances in English scheduling correspond to some 30000 sentences.

Integrating Different Learning Approaches into a Multilingual Spoken Language Translation System*

P. Geutner[1], B. Suhm[2], F.-D. Buø[1], T. Kemp[1], L. Mayfield[2], A. E. McNair[1], I. Rogina[1], T. Schultz[1], T. Sloboda[1], W. Ward[2], M. Woszczyna[1] and A. Waibel [1,2]

pgeutner@ira.uka.de

Interactive Systems Laboratories
[1] Karlsruhe University (Germany)
[2] Carnegie Mellon University (USA)

Abstract

Building multilingual spoken language translation systems requires knowledge about both acoustic models and language models of each language to be translated. Our multilingual translation system JANUS-2 is able to translate English and German spoken input into either English, German, Spanish, Japanese or Korean output. Getting optimal acoustic and language models as well as developing adequate dictionaries for all these languages requires a lot of hand-tuning and is time-consuming and labor intensive. In this paper we will present learning techniques that improve acoustic models by automatically adapting codebook sizes, a learning algorithm that increases and adapts phonetic dictionaries for the recognition process and also a statistically based language model with some linguistic knowledge that increases recognition performance. To ensure a robust translation system, semantic rather than syntactic analysis is done. Concept based speech translation and a connectionist parser that learns to parse into feature structures are introduced. Furthermore, different repair mechanisms to recover from recognition errors will be described.

*Our German recognition engine, developed at the University of Karlsruhe, is part of the VERBMOBIL project and VERBMOBIL systems developed under BMBF funding. The Spanish speech translation module has been developed at Carnegie Mellon University under project ENTHUSIAST funded by the US Government. Other components are under development in collaboration with partners of the C-STAR Consortium.

Introduction

Our multilingual spoken language translation system JANUS-2 (Woszczyna et al. 1993; Suhm et al. 1994b) evolved from the previous JANUS (Osterholtz 1992) system which was able to process syntactically well-formed read speech within a certain domain and a limited vocabulary of 500 words. JANUS-2 processes spontaneous human-to-human dialogs in a scheduling domain where the vocabulary – depending on the language – may vary between 2000 and 3000 words. The JANUS-2 system provides a framework under which complementary speech translation system components from different projects, like VERBMOBIL (Wahlster 1993), ENTHUSIAST and the C-STAR Consortium (C-STAR 1994), can be integrated and compared. Currently, English and German spoken input can be translated into either English, German, Spanish, Japanese or Korean output. Work is in progress to add Spanish and Korean as input languages.

We will propose a data-driven learning approach for automatic codebook adaptation based on amount and distribution of data to improve the acoustic models within the speech recognizer. Second, a method for automatically increasing and adapting a phonetic dictionary will be introduced. Moreover, a statistically based approach is combined with linguistic knowledge to create morpheme-based language models. Also a new approach towards robust translation of spoken language will be presented. We briefly describe a parsing and translation approach based on an interlingua text (ILT), where an interlingua is intended to be a language-independent representation of meaning. Besides, the functionality of a connectionist parser that learns to parse into feature structures is shown. Fi-