View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by KITopen

Expanding the Domain of a Multi-lingual Speech-to-Speech Translation System

Alon Lavie, Lori Levin, Puming Zhan, Maite Taboada, Donna Gates, Mirella Lapata, Cortis Clark, Matthew Broadhead, Alex Waibel
Interactive Systems Laboratory
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
email : lavie@cs.cmu.edu

Abstract

JANUS is a multi-lingual speech-to-speech translation system, which has been designed to translate spontaneous spoken language in a limited domain. In this paper, we describe our recent preliminary efforts to expand the domain of coverage of the system from the rather limited Appointment Scheduling domain, to the much richer Travel Planning domain. We compare the two domains in terms of out-of-vocabulary rates and linguistic complexity. We discuss the challenges that these differences impose on our translation system and some planned changes in the design of the system. Initial evaluations on Travel Planning data are also presented.

Introduction

Spoken language understanding systems have been reasonably successful in limited semantic domains[1]. The limited domains naturally constrain vocabulary and perplexity, making speech recognition tractable. In addition, the relatively small range of meanings that could be conveyed make parsing and understanding tractable. Now, with the increasing success of large vocabulary continuous speech recognition (LVCSR), the challenge is to similarly scale up spoken language understanding. In this paper we describe our plans for extending the JANUS speech-to-speech translation system [1] [2] from the Appointment Scheduling domain to a broader domain, Travel Planning, which has a rich sub-domain structure, covering many topics.

In the last three years, the JANUS project has been developing a speech-to-speech translation system for the Appointment Scheduling domain (two people setting up a time to meet with each other). Although the data we have been working with is spontaneous speech, the scheduling scenario naturally limits the vocabulary to about 3000 words in English and about 4000 words in Spanish and German, which have more inflection. Similarly, the types of dialogues are naturally limited. A scheduling dialogue typically consists of opening greetings, followed by several rounds of negotiation on a time, followed by closings. There is ambiguity, for example whether a number refers to a date or a time, but many potentially ambiguous sentences have only one possible meaning in the scheduling domain. To date, our translation system for the scheduling domain has achieved performance levels on unseen data of over 80% acceptable translations on transcribed input, and over 70% acceptable translations on speech input recognized with a 75-90% word accuracy, depending on the language.

In addition to the scheduling domain, the JANUS speech recognizer has also been trained and developed for Switchboard, a broad domain LVCSR task. We are now planning to expand our domain of spoken language understanding as well. The new domain, Travel Planning, is still limited, but is significantly more complex than the scheduling domain. Travel Planning contains a number of semantic sub-domains — for example, accommodation, events, transportation — each of which has a number of sub-topics such as time, location, and price. Travel planning also differs from scheduling in having more types of interactions. Scheduling consists almost entirely of negotiation dialogues except for openings and closings. The travel domain includes negotiations, information seeking, instruction giving, and dialogues that accompany non-linguistic domain actions such as paying and reserving. Furthermore, there is more ambiguity in travel planning, especially because the same utterance can have different meanings in different sub-domains.

An important part of our approach to the travel planning domain is a system of sub-domain parsing. Each sentence will be parsed in parallel by a number of sub-domain grammars, each of which is faster and less ambiguous than a large grammar would be. Since the sub-grammars are separated from each other, the ambiguities between them will add and not multiply. The content of each sub-domain grammar will be determined automatically by running a comprehensive grammar over a corpus in which each sentence has a sub-domain tag.

[1] Verbmobil, systems developed under the ATIS initiative, and systems developed at SRI, AT&T and MIT/Lincoln Lab are examples of such successful spoken language understanding systems.

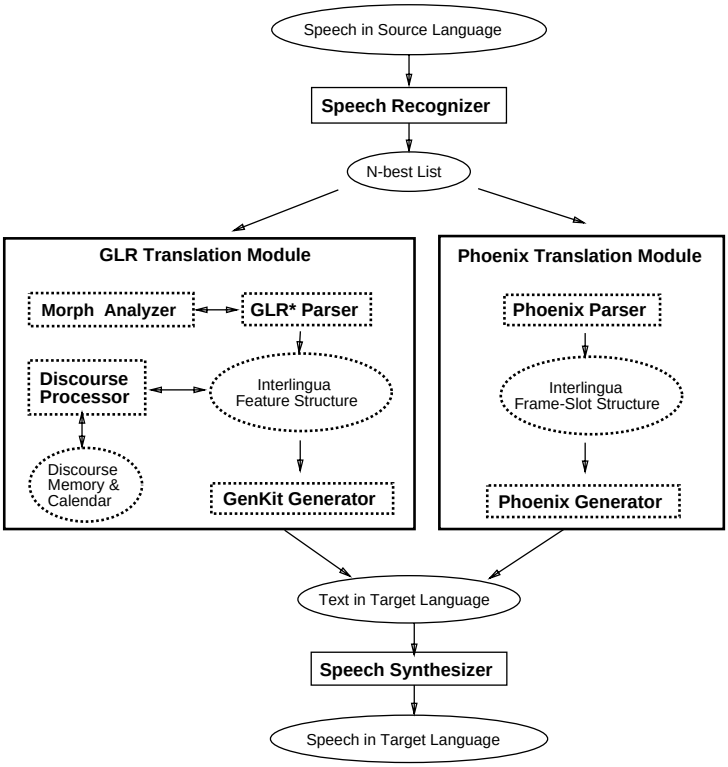

Figure 1: **The JANUS System**

In the remaining sections, we summarize the JANUS approach to spoken language translation, highlight the differences between the scheduling and travel planning domains, present some preliminary results for the travel planning domain, and summarize our plans for modifying the design of the system, in order to effectively handle a variety of sub-domains.

Review of our approach

A component diagram of our system for the Scheduling domain can be seen in Figure 1. The main system modules are speech recognition, parsing, discourse processing, and generation. Each module is language independent in the sense that it consists of a general processor that can be loaded with language specific knowledge sources. The translation system is based on an interlingua approach. The source language input string is first analyzed by a parser, which produces a language-independent interlingua content representation. The interlingua is then passed to a generation component, which produces an output string in the target language. In an attempt to achieve both robustness and translation accuracy when faced with speech disfluencies and recognition errors, we use two different parsing strategies: a GLR parser designed to be more accurate, and a Phoenix parser designed to be more robust. Detailed descriptions of the system components appear in our previous publications [1] [2] [3] [4] [5] [6].

Speech translation in the JANUS system is guided by the general principle that spoken utterances can be analyzed and translated as a sequential collection of *semantic dialogue units* (SDUs), each of which roughly corresponds to a speech-act. SDUs are semantically coherent pieces of information. The interlingua representation in our system was designed to capture meaning at the level of such SDUs. Each semantic dialogue unit is analyzed into an interlingua representation.

For both parsers, segmentation of an input utterance into SDUs is achieved in a two-stage process, partly prior to and partly during parsing. Pre-parsing segmentation relies on acoustic, lexical, syntactic, semantic, and statistical knowledge sources. We use a statistical measure that attempts to capture the likelihood of an SDU boundary between any two words of an utterance. The measure is trained on hand-segmented transcriptions of dialogues. Pre-parsing segmentation substantially reduces parsing time, increases parse accuracy, and reduces ambiguity. Final segmentation into SDUs is done during parse time, guided by the grammar rules. The same statistical measure used to find the most likely SDU boundaries during pre-parsing segmentation is used to filter out unlikely segmentations during parse time.

For the scheduling domain, we have been using semantic grammars, in which the grammar rules define semantic categories such as `busy-free-phrase` and

`schedule-meeting` in addition to syntactic categories such as `NP` and `VP`. There were several reasons for chosing semantic grammars. First, the domain lends itself well to semantic grammars because there are many fixed expressions and common expressions that are almost formulaic. Breaking these down syntactically would be an unnecessary complication. Additionally, spontaneous spoken language is often syntactically ill formed, yet semantically coherent. Semantic grammars allow our robust parsers to extract the key concepts being conveyed, even when the input is not completely grammatical in a syntactic sense. Furthermore, we wanted to achieve reasonable coverage of the domain in as short a time as possible. Our experience has been that, for limited domains, 60% to 80% coverage can be achieved in a few months with semantic grammars.

In order to assess the overall effectiveness of the translation system, we developed a detailed end-to-end evaluation procedure [7]. We evaluate the translation modules on both transcribed and speech recognized input. The evaluation of transcribed input allows us to assess how well our translation modules would function with "perfect" speech recognition. Testing is performed on a set of unseen dialogues that were not used for developing the translation modules or training the speech recognizer.

The translation of an utterance is manually evaluated by assigning it a grade or a set of grades based on the number of SDUs in the utterance. Each SDU is classified first as either relevant to the scheduling domain (in-domain) or not relevant to the scheduling domain (out-of-domain). Each SDU is then assigned one of four grades for translation quality: (1) Perfect - a fluent translation with all information conveyed; (2) OK - all important information translated correctly but some unimportant details missing, or the translation is awkward; (3) Bad - unacceptable translation; (4) Recognition Error - unacceptable translation due to a speech recognition error. These grades are used for both in-domain and out-of-domain sentences. However, if an out-of-domain sentence is automatically detected as such by the parser and is not translated at all, it is given an "OK" grade. The evaluations are performed by one or more independent graders. When more than one grader is used, the results are averaged together.

Comparison of Travel and Scheduling Domains

In this section we compare some characteristics of the English Travel Domain (ETD) and the English Spontaneous Scheduling Task (ESST). The ETD and ESST databases are not comparable in some ways — ETD has been under development for less than one year whereas the ESST database was collected over a three year period and is much larger. Also, the ESST recording scenario was push-to-talk whereas the ETD recording setup allows for cross talk. However, it is possible to draw some comparisons. For example, speech recognition appears to indicate that the ETD domain has a higher out-of-vocabulary rate. In addition, informal observations of the grammar developers point out sources of ambiguity in ETD that do not exist in ESST.

ESST data was collected by giving marked-up calendars to two speakers and asking them to schedule a two hour meeting at a time that was free on each of their calendars. This method allowed us to collect speech in a limited domain that was nevertheless spontaneous. Similarly, ETD data is collected in a simulated conversation between a traveller and a travel agent. The speaker playing the traveller is given a scenario such as "You are travelling with your wife and teenage daughter to the Pittsburgh Arts Festival. Book a hotel room that is conveniently located." The speaker playing the travel agent has information about hotels, transportation, etc. on which to base answers to the traveller's questions.

The current ETD database contains 2000 utterances (30 dialogues). For both speech recognition and grammar development, we used 1292 utterances (20 dialogues) as a training set and 368 utterances (5 dialogues) as a test set. The ESST speech recognition training set contains over 40 hours speech data and is composed of 8277 utterances. The testing set is composed of 612 utterances. The ESST testing vocabulary contains 2900 words. The current word error rate of the ESST recognizer is about 23%.

Some differences in the ETD and ESST databases are attributable to the push-to-talk vs. cross-talk recording scenarios. In push-to-talk dialogues, the participants push a key when they start and finish speaking, and cannot speak at the same time. In cross-talk dialogues, participants can speak freely and their speech can overlap. The average length of ESST push-to-talk utterances is 33.6 words. ETD cross-talk utterances average 14.6 words. In addition, the noise rate (noise-tokens/total-tokens) is 25.3% for the ESST training set, and 15.23% for the travel domain training set.

In spite of the differences in the size of the two databases, we can compare the out-of-vocabulary rates in order to get some idea of the difference in vocabulary sizes of the two domains. The vocabulary size of the ESST system is 2900 words, which includes all unique words in the ESST training set. The ETD speech vocabulary was constructed by augmenting the ESST vocabulary with 312 new words that appeared in the ETD training set. This results in a vocabulary of 3212 words. The ETD test set contains 272 out-of-vocabulary tokens out of a total of 2554 tokens. Thus, the out-of-vocabulary rate for the ETD test set is 10.65%. This compares with out-of-vocabulary rates for ESST that have ranged between 1% to 4%. We have also found noticeable language model perplexity differences between the ESST and ETD domains. However, these appear to be highly dependent on the method used for obtaining the language models, and did not seem to form a consistent pattern.

There are also differences between ETD and ESST with respect to parsing and ambiguity. For example, in the scheduling domain, numbers could be either dates or times. In the travel domain, a number like *twelve fifteen* could be a time, price (twelve dollars and fifteen cents or one thousand two hundred and fifteen dollars), room number, flight number, etc. The increase in interpretations can be attributed to the larger number of sub-domains.

Preliminary Results for the Travel Planning Domain

Speech Recognition

Due to the very limited amount of training data available for the travel domain, we decided to attempt to build a speech recognition system for ETD by a process of adapting the acoustic and language models of our ESST recognition system. To start off, we conducted a preliminary evaluation on the ETD test set using the original ESST acoustic and language models. With this set-up, the average word error rate on the ETD test set was 55%. Next, we added the ETD training corpus to the ESST training corpus and used the merged corpus for language model training. With this new language model, we obtained a 42% word error rate. We also tried to build the language model just based on the ETD corpus, which was smoothed by interpolation with the ESST language model. However, this resulted in only about 0.5% improvement.

In the next stage, to allow for better training with very limited amounts of data, we rebuilt the acoustic models using just the PLP feature and signal energy. This dramatically reduced the codebook size and the dimension of the feature vectors. With the new acoustic models which were trained with ESST and ETD speech data, we obtained a 37.5% word error rate. Training the acoustic models with Vocal Tract Normalization (VTLN) speaker normalization reduced the word error rate even further to 35.8%. We experimented with adapting the ESST acoustic models by using the ETD speech as adaptation data, but both the MLLR and MAP adaptation methods did not reduce the word error rate any further.

There are three main reasons why the word error rate is much higher for ETD than ESST. First, the out-of-vocabulary rate is significantly higher. Second, because the travel domain database is very small compared to the ESST database, the ESST data dominates the acoustic and language models. Third, the ETD data is cross-talk, which is generally more disfluent and contains more co-articulation. (This was demonstrated with our Spanish Spontaneous Scheduling Task database, which contained both push-to-talk and cross-talk utterances.) We expect significantly larger amounts of training data to at least partially alleviate these problems resulting in significant performance gains.

We obtained the above results without using the ETD speech data to train the acoustic models. Considering that the travel speech data is only a very small portion of all the available English training data, we plan to use adaptation techniques to adapt the current ESST acoustic models into models for the travel domain.

Translation Components

In addition to speech recognition, we have done some preliminary development of our translation components for ETD. Since we currently have only English travel data, we developed English analysis and generation grammars for English-to-English translation (or paraphrase) using the Phoenix system. On a test set of six unseen dialogues, we achieve about 45% acceptable translation of transcribed SDUs in the travel domain.[2]

A preliminary interlingua design for the travel domain contains about 200 concepts arranged in an IS-A hierarchy, semantic features to represent the meaning of closed class items, and a list of five basic speech acts which each have several sub-types. We have developed experimental grammars that are compatible with the interlingua design for English parsing (Phoenix), English generation (Phoenix and GLR), German generation (Phoenix), and Japanese generation (Phoenix). Mappers mediate between Phoenix tree structures and the feature structures of the interlingua design.

Planned Modifications to the System Design

We believe that the main challenge that the Travel Planning domain will impose on our translation system is the problem of how to effectively deal with significantly greater levels of ambiguity. We suspect that the single semantic grammar approach, which we have been following for the scheduling domain, will not be feasible for the Travel domain. Syntactically similar structures that correspond to different semantic concepts usually require separate rules in a semantic grammar. Thus, as the domain semantically expands, the size of the semantic grammar tends to substantially grow. With this growth, significant new ambiguities are introduced into the grammar, and these tend to multiply.

One method of dealing with this problem is by "breaking" the large travel domain into several semantic sub-domains. Because each of the sub-domains will be semantically much more narrow, the corresponding semantic grammars should be smaller and far less ambiguous, leading to faster parsing and more accurate analysis. Since the sub-grammars are separated from each other, the ambiguities between them will add and not multiply.

[2]The travel domain grammars have been under development for only a few months. The scheduling domain grammars, which have been under development for three years achieve about 85% acceptable translations on unseen transcribed input.

Travel domain dialogues, however, will often contain sub-dialogues and utterances from different sub-domains, and will likely shift between one sub-domain and another. We thus envision modifying the design of our translation system to facilitate dealing with multiple sub-domains simultaneously and/or in parallel. Utterances will be first segmented into sub-utterances by a segmentation procedure. We expect that in most cases, each sub-utterance will not span multiple sub-domains. Each sub-utterance will then be parsed in parallel by a number of sub-domain grammars, each of which is faster and less ambiguous than a large grammar would be. Because each sub-domain grammar should be able to parse well only sentences that fall in its domain of coverage, we expect that in many cases it should be relatively easy to select which among the parses produced by the different sub-domain grammars is most appropriate and/or correct. Sentences that are covered well by more than one grammar most likely indicate true semantic ambiguity (for example, as mentioned above, an expression such as `twelve fifteen`, which can be interpreted as a time, flight number, room number or price). To aid in such cases, we plan on developing a sub-domain/topic identification and tracking component that will be independent of the semantic grammars. This component will assist in disambiguating among semantically ambiguous analyses using contextual information, modeled via statistical and other methods.

The effectiveness of the sub-domain approach described above will most likely depend heavily on our ability to choose appropriate sub-domains. Sub-domains should be chosen to be semantically distinct, so that sentences may be easily classified into sub-domains by both humans and machine. Our current sub-domain classification has two dimensions. The first distinguishes between topics such as accommodation, transportation, restaurants, events and sights. The second distinguishes between discussions about price, reservations, location, time, participants, directions and general information. We are in the process of experimenting with both possible classifications, and their combinations. We have constructed a simple sub-domain classifier that is based on a naive-Bayesian approach and trained on the available ETD data. Preliminary tests (on unseen data) indicate that the simple classifier correctly identifies sub-domains classified according to the first dimension about 65% of the time. When the second dimension set of sub-domain classifications is used, the classifier correctly identifies 75% of the sub-domains.

We would like to avoid having to manually construct the different sub-domain grammars for several reasons. First, even if the various sub-domains are semantically distinct, multiple sub-domain grammars will likely contain some of the same rules. Furthermore, since we expect to experiment with various sub-domain classifications, it would be useful to devise an automatic method for dividing a large comprehensive grammar of the entire travel domain into sub-domain grammars. We plan to achieve this task by running a comprehensive grammar over a corpus in which each sentence is tagged with its corresponding sub-domain and correct parse. The grammar rules that correspond to the correct parse are then added to the appropriate sub-domain grammar. This approach is similar to one proposed by Rayner and Samuelsson [8] for tailoring a large grammar to a given corpus.

Conclusions

In this paper we described our plans for extending the JANUS speech-to-speech translation system from the Appointment Scheduling domain to a broader domain, Travel Planning, which has a rich sub-domain structure. Our preliminary experiments with English travel domain data indicate that it is characterized by higher out-of-vocabulary rates and greater levels of semantic complexity, compared with English scheduling domain data. In order to effectively deal with the significantly greater levels of ambiguity, we plan to use a collection of sub-domain grammars, which will in sum cover the entire travel planning domain. Our system design will be modified to facilitate working with multiple sub-domain grammars in parallel. The collection of appropriate sub-domains will be determined empirically. Automatic pruning methods will be used to derive each of the sub-domain grammars from a manually constructed comprehensive grammar. We expect to complete an initial prototype implementation of the above methods and have additional preliminary evaluations of their effectiveness by late summer 1997.

Acknowledgements

The work reported in this paper was funded in part by grants from ATR - Interpreting Telecommunications Research Laboratories of Japan, the US Department of Defense, and the Verbmobil Project of the Federal Republic of Germany.

References

[1] A. Lavie, D. Gates, M. Gavaldá, L. Mayfield, A. Waibel, and L. Levin. Multi-lingual translation of spontaneously spoken language in a limited domain. In *Proceedings of the COLING*, 1996.

[2] Alon Lavie, Alex Waibel, Lori Levin, Donna Gates, Marsal Gavalda, Torsten Zeppenfeld, Puming Zhan and Oren Glickman. Translation of Conversational Speech with JANUS-II, In Proceedings of ICSLP-96, Philadelphia, USA, October 1996.

[3] P. Zhan, K. Ries, M. Gavalda, D. Gates, A. Lavie and A. Waibel. *JANUS-II: Towards Spontaneous Spanish Speech Recognition,* Proceedings of ICSLP-96, Philadelphia, PA, October 1996

[4] A. Lavie. *A Grammar Based Robust Parser For Spontaneous Speech.* PhD thesis, School of Computer Science, Carnegie Mellon University, 1995.

[5] L. Mayfield, M. Gavaldà, Y-H. Seo, B. Suhm, W. Ward, A. Waibel. *Parsing Real Input in JANUS: a Concept-Based Approach,* In Proceedings of TMI 95.

[6] Y. Qu, C. P. Rose, B. Di Eugenio. Using Discourse Predictions for Ambiguity Resolution, In Proceedings of COLING-96, Copenhagen, Denmark, August 1996.

[7] D. Gates, A. Lavie, L. Levin, A. Waibel, M. Gavaldà, L. Mayfield, M. Woszczyna and P. Zhan. *End-to-end Evaluation in JANUS: a Speech-to-speech Translation System,* To appear in Proceedings of ECAI Workshop on Dialogue Processing in Spoken Language Systems, Budapest, Hungary, August 1996.

[8] M-S. Agnäs, H. Alshawi, I. Bretan, D. Carter, K. Ceder, M. Collins, R. Crouch, V. Digalakis, B. Ekholm, B. Gamback, J. Kaja, J. Karlgren, B. Lyberg, P. Price, S. Pulman, M. Rayner, C. Samuelsson, and T. Svensson. Spoken Language Translator: First-Year Report. Technical Report CRC-043, SRI Cambridge, 1994.