Domain Portability in Speech-to-Speech Translation

Alon Lavie, Lori Levin, Tanja Schultz, Chad Langley, Benjamin Han
Alicia Tribble, Donna Gates, Dorcas Wallace and Kay Peterson
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
alavie@cs.cmu.edu

1. INTRODUCTION

Speech-to-speech translation has made significant advances over the past decade, with several high-visibility projects (C-STAR, Verbmobil, the Spoken Language Translator, and others) significantly advancing the state-of-the-art. While speech recognition can currently effectively deal with very large vocabularies and is fairly speaker independent, speech translation is currently still effective only in limited, albeit large, domains. The issue of domain portability is thus of significant importance, with several current research efforts designed to develop speech-translation systems that can be ported to new domains with significantly less time and effort than is currently possible.

This paper reports on three experiments on portability of a speech-to-speech translation system between semantic domains.[1] The experiments were conducted with the JANUS system [5, 8, 12], initially developed for a narrow travel planning domain, and ported to the doctor-patient domain and an extended tourism domain. The experiments cover both rule-based and statistical methods, and hand-written as well as automatically learned rules. For rule-based systems, we have investigated the re-usability of rules and other knowledge sources from other domains. For statistical methods, we have investigated how much additional training data is needed for each new domain. We are also experimenting with combinations of hand-written and automatically learned components. For speech recognition, we have conducted studies of what parameters change when a recognizer is ported from one domain to another, and how these changes affect recognition performance.

2. DESCRIPTION OF THE INTERLINGUA

The first two experiments concern the analysis component of our interlingua-based MT system. The analysis component takes a sentence as input and produces an interlingua representation as output. We use a task-oriented interlingua [4, 3] based on domain actions. Examples of domain actions are giving information about the onset of a symptom (e.g., *I have a headache*) or asking a patient to perform some action (e.g., *wiggle your fingers*). The interlingua, shown in the example below, has five main components: (1) a speaker tag such as `a:` for doctor (agent) and `c:` for a patient (customer), (2) a speech act, in this case, `give-information` (3) some concepts (`+body-state` and `+existence`), and (4) some arguments (`body-state-spec=` and `body-location=`), and (5) some sub-arguments (`identifiability=no` and `inside=head`).

```
I have a pain in my head.
c:give-information+existence+body-state
   (body-state-spec=(pain,identifiability=no),
    body-location=(inside=head))
```

[1]We have also worked on the issue of portability across languages via our interlingua approach to translation [3] and on portability of speech recognition across languages [10].

3. EXPERIMENT 1: EXTENSION OF SEMANTIC GRAMMAR RULES BY HAND AND BY AUTOMATIC LEARNING

Experiment 1 concerns extension of the coverage of semantic grammars in the medical domain. Semantic grammars are based on semantic constituents such as request information phrases (e.g., *I was wondering ...*) and location phrases (e.g., *in my right arm*) rather than syntactic constituents such as noun phrases and verb phrases. In other papers [12, 5], we have described how our modular grammar design enhances portability across domains. The portable grammar modules are the cross-domain module, containing rules for things like greetings, and the shared module, containing rules for things like times, dates, and locations. Figure 1 shows a parse tree for the sentence *How long have you had this pain?* `XDM` indicates nodes that were produced by cross-domain rules. `MED` indicates nodes that were produced by rules from the new medical domain grammar.

The preliminary doctor-patient grammar focuses on three medical situations: `give-information+existence` — giving information about the existence of a symptom (*I have been getting headaches*); `give-information+onset` – giving information about the onset of a symptom (*The headaches started three months ago*); and `give-information+occurrence` — giving information about the onset of an instance of the symptoms (*The headaches start behind my ears*). Symptoms are expressed as `body-state` (e.g., pain), `body-object` (e.g., rash), and `body-event` (e.g., bleeding).

Our experiment on extendibility was based on a hand written seed grammar that was extended by hand and by automatic learning. The seed grammar covered the domain actions mentioned above, but did not cover very many ways to phrase each domain action. For example, it might have covered *The headaches started*

```
[request-information+existence+body-state]::MED
(  WH-PHRASES::XDM
    ( [q:duration=]::XDM ( [dur:question]::XDM ( how long ) ) )
   HAVE-GET-FEEL::MED ( GET ( have ) ) you
   HAVE-GET-FEEL::MED ( HAS ( had ) )
   [super_body-state-spec=]::MED
    ( [body-state-spec=]::MED
        ( ID-WHOSE::MED
           ( [identifiability=]
             ( [id:non-distant] ( this ) ) )
          BODY-STATE::MED ( [pain]::MED ( pain ) ) ) ) )
```

Figure 1: Parser output with nodes produced by medical and cross-domain grammars.

	Seed	Extended	Learned
IF	37.2	37.2	31.3
Domain Action	37.2	37.2	31.3
Speech Act			
Recall	43.3	48.2	49.3
Precision	71.0	75.0	45.8
Concept List			
Recall	2.2	10.1	32.5
Precision	12.5	42.2	25.1
Top-Level Arguments			
Recall	0.0	7.2	29.6
Precision	0.0	42.2	34.4
Top-Level Values			
Recall	0.0	8.3	29.8
Precision	0.0	50.0	39.2
Sub-Level Arguments			
Recall	0.0	28.3	14.1
Precision	0.0	48.2	12.6
Sub-level Values			
Recall	1.2	28.3	14.1
Precision	6.2	48.2	12.9

Table 1: Comparison of seed grammar, human-extended grammar, and machine-learned grammar on unseen data

three months ago but not *I started getting the headaches three months ago*. The seed grammar was extended by hand and by automatic learning to cover a development set of 133 utterances. The result was two new grammars, a human-extended grammar and a machine-learned grammar, referred to as the extended and learned grammars in Table 1. The two new grammars were then tested on 132 unseen sentences in order to compare generality of the rules. Results are reported only for 83 of the 132 sentences which were covered by the current interlingua design. The remaining 49 sentences were not covered by the current interlingua design and were not scored. Results are shown in Table 1.

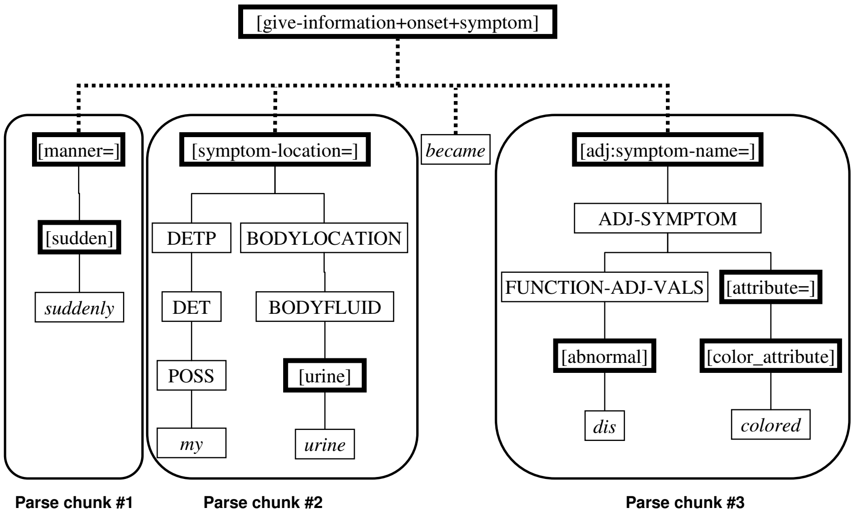
The parsed test sentences were scored in comparison to a hand-coded correct interlingua representation. Table 1 separates results for six components of the interlingua: speech act, concepts, top-level arguments, top-level values, sub-level arguments, and sub-level values, in addition to the total interlingua, and the domain action (speech act and concepts combined). The components of the interlingua were described in Section 2.

The scores for the total interlingua and domain action are reported as percent correct. The scores for the six components of the interlingua are reported as average percent precision and recall. For example, if the correct interlingua for a sentence has two concepts, and the parser produces three, two of which are correct and one of which is incorrect, the precision is 66% and the recall is 100%.

Several trends are reflected in the results. Both the human-extended grammar and the machine-learned grammar show improved performance over the seed grammar. However, the human extended grammar tended to outperform the automatically learned grammar in precision, whereas the automatically learned grammar tended to outperform the human extended grammar in recall. This result is to be expected: humans are capable of formulating correct rules, but may not have time to analyze the amount of data that a machine can analyze. (The time spent on the human extended grammar after the seed grammar was complete was only five days.)

Grammar Induction: Our work on automatic grammar induction for Experiment 1 is still in preliminary stages. At this point, we have experimented with completely automatic induction (no interaction with a user)[2] of new grammar rules starting from a core grammar and using a development set of sentences that are not parsable according to the core grammar. The development sentences are tagged with the correct interlingua, and they do not stray from the concepts covered by the core grammar — they only correspond to alternative (previously unseen) ways of expressing the same set of covered concepts. The automatic induction is based on performing tree matching between a skeletal tree representation obtained from the interlingua, and a collection of parse fragments

[2] Previous work on our project [2] investigated learning of grammar rules with user interaction.




```
Original interlingua:
give-information+onset+symptom
(symptom-name=(abnormal,attribute=color_attribute),symptom-location=urine,
manner=sudden)

Learned Grammar Rule:
s[give-information+onset+symptom]
( [manner=] [symptom-location=] *+became [adj:symptom-name=] )
```

Figure 2: A reconstructed parse tree from the Interlingua

that is derived from parsing the new sentence with the core grammar. Extensions to the existing rules are hypothesized in a way that would produce the correct interlingua representation for the input utterance.

Figure 2 shows a tree corresponding to an automatically learned rule. The input to the learning algorithm is the interlingua (shown in bold boxes in the figure) and three parse chunks (circled in the figure). The dashed edges are augmented by the learning algorithm.

4. EXPERIMENT 2: PORTING TO A NEW DOMAIN USING A HYBRID RULE-BASED AND STATISTICAL ANALYSIS APPROACH

We are in the process of developing a new alternative analysis approach for our interlingua-based speech-translation systems that combines rule-based and statistical methods and we believe inherently supports faster porting into new domains. The main aspects of the approach are the following. Rather than developing complete semantic grammars for analyzing utterances into our interlingua (either completely manually, or using grammar induction techniques), we separate the task into two main levels. We continue to develop and maintain rule-based grammars for phrases that correspond to argument-level concepts of our interlingua representation (e.g., time expressions, locations, symptom-names, etc.). However, instead of developing grammar rules for assembling the argument-level phrases into appropriate domain actions, we apply machine learning and classification techniques [1] to learn these mappings from a corpus of interlingua tagged utterances. (Earlier work on this task is reported in [6].)

We believe this approach should prove to be more suitable for fast porting into new domains for the following reasons. Many of the required argument-level phrase grammars for a new domain are likely to be covered by already existing grammar modules, as can be seen by examining the XDM (cross-domain) nodes in Figure 1. The remaining new phrase grammars are fairly fast and straightforward to develop. The central questions, however, are whether the statistical methods used for classifying strings of arguments into domain actions are accurate enough, and what amounts of tagged data are required to obtain reasonable levels of performance. To assess this last question, we tested the performance of the current speech-act and concept classifiers for the expanded travel-domain when trained with increasing amounts of training data. The results of these experiments are shown in Figure 3. We also report the performance of the domain-action classification derived from the combined speech-act and concepts. As can be seen, performance reaches a relative plateau at around 4000-5000 utterances. We see these results as indicative that this approach should indeed prove to be significantly easier to port to new domains. Creating a tagged database of this order of magnitude can be done in a few weeks, rather than the months required for complete manual grammar development time.

5. EXPERIMENT 3: PORTING THE SPEECH RECOGNIZER TO NEW DOMAINS

When the speech recognition components (acoustic models, pronunciation dictionary, vocabulary, and language model) are ported across domains and languages mainly three types of mismatches

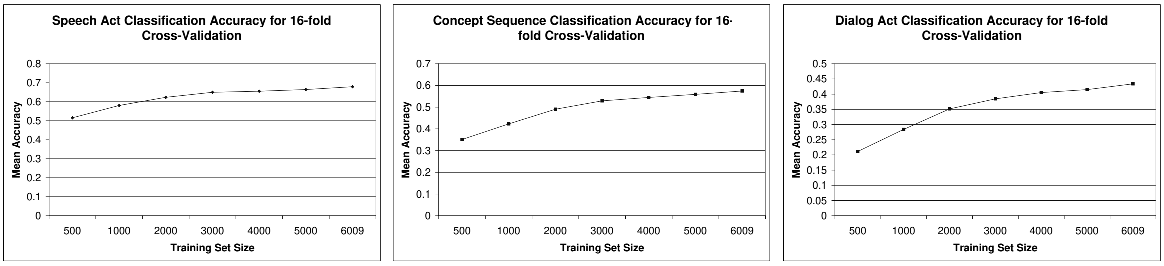


Figure 3: Performance of Speech-Act, Concept, and Domain-Action Classifiers Using Increasing Amounts of Training Data

Baseline Systems WER on Different Tasks [%]	
BN (Broadcast News) h4e98_1, all F-conditions	18.5
ESST (scheduling and travel planning domain)	24.3
BN+ESST	18.4
C-STAR (travel planning domain)	20.2
Adaptation → Meeting Recognition	
ESST on meeting data	54.1
BN on meeting data	44.2
+ acoustic MAP Adaptation (10h meeting data)	40.4
+ language model interpolation (16 meetings)	38.7
BN+ESST on meeting data	42.2
+ language model interpolation (16 meetings)	39.0
Adaptation → Doctor-Patient Domain	
C-STAR on doctor-patient data	34.1
+ language model interpolation (≈ 34 dialogs)	25.1

Table 2: Recognition Results

occur: (1) mismatches in recording condition; (2) speaking style mismatches; as well as (3) vocabulary and language model mismatches. In the past these problems have mostly been solved by collecting large amounts of acoustic data for training the acoustic models and development of the pronunciation dictionary, as well as large text data for vocabulary coverage and language model calculation. However, especially for highly specialized domains and conversational speaking styles, large databases cannot always be provided. Therefore, our research has focused on the problem of how to build LVCSR systems for new tasks and languages [7, 9] using only a limited amount of data. In this third experiment we investigate the results of porting the speech recognition component of our MT system to different new domains. The experiments and improvements were conducted with the Janus Speech Recognition Toolkit JRTk [13].

Table 2 shows the results of porting four baseline speech recognition systems to the doctor-patient domain, and to the meeting domain. The four baseline systems are trained on Broadcast News (BN), English Spontaneous Scheduling Task (ESST), combined BN and ESST, and the travel planning domain of the C-STAR consortium (`http://www.c-star.org`). The given tasks illustrate a variety of domain size, speaking styles and recording conditions ranging from clean spontaneous speech in a very limited domain (ESST, C-STAR) to highly conversational multi-party speech in an extremely broad domain (Meeting). As a consequence the error rates on the meeting data are quite high but using MAP (Maximum A Posteriori) acoustic model adaptation and language model adaptation the error rate can be reduced by about 10.2% relative over the BN baseline system. With the doctor-patient data the drop in error rate was less severe which can be explained by the similar speaking style and recording conditions for C-STAR and doctor-patient data. Details about the applied recognition engine can be found in [10] for ESST and [11] for the BN system.

6. ACKNOWLEDGMENTS

The research work reported here was funded in part by the DARPA TIDES Program and supported in part by the National Science Foundation under Grant number 9982227. Any opinions, findings and conclusions or recomendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF) or DARPA.

7. REFERENCES

[1] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. TiMBL: Tilburg Memory Based Learner, version 3.0 Reference Guide. Technical Report Technical Report 00-01, ILK, 2000. Avaliable at http://ilk.kub.nl/ ilk/papers/ilk0001.ps.gz.

[2] M. Gavaldà. Epiphenomenal Grammar Acquisition with GSG. In *Proceedings of the Workshop on Conversational Systems of the 6th Conference on Applied Natural Language Processing and the 1st Conference of the North American Chapter of the Association for Computational Linguistics (ANLP/NAACL-2000)*, Seattle, U.S.A, May 2000.

[3] L. Levin, D. Gates, A. Lavie, F. Pianesi, D. Wallace, T. Watanabe, and M. Woszczyna. Evaluation of a Practical Interlingua for Task-Oriented Dialogue. In *Workshop on*

Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP, Seattle, 2000.

[4] L. Levin, D. Gates, A. Lavie, and A. Waibel. An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'98)*, pages Vol. 4, 1155–1158, Sydney, Australia, 1998.

[5] L. Levin, A. Lavie, M. Woszczyna, D. Gates, M. Gavaldà, D. Koll, and A. Waibel. The Janus-III Translation System. *Machine Translation*. To appear.

[6] M. Munk. Shallow statistical parsing for machine translation. Master's thesis, University of Karlsruhe, Karlsruhe, Germany, 1999. `http://www.is.cs.cmu.edu/papers/speech/masters-thesis/MS99.munk.ps.gz`.

[7] T. Schultz and A. Waibel. Polyphone Decision Tree Specialization for Language Adaptation. In *Proceedings of the ICASSP*, Istanbul, Turkey, 2000.

[8] A. Waibel. Interactive Translation of Conversational Speech. *Computer*, 19(7):41–48, 1996.

[9] A. Waibel, P. Geutner, L. Mayfield-Tomokiyo, T. Schultz, and M. Woszczyna. Multilinguality in Speech and Spoken Language Systems. *Proceedings of the IEEE, Special Issue on Spoken Language Processing*, 88(8):1297–1313, 2000.

[10] A. Waibel, H. Soltau, T. Schultz, T. Schaaf, and F. Metze. *Multilingual Speech Recognition*, chapter From Speech Input to Augmented Word Lattices, pages 33–45. Springer Verlag, Berlin, Heidelberg, New York, artificial Intelligence edition, 2000.

[11] A. Waibel, H. Yu, H. Soltau, T. Schultz, T. Schaaf, Y. Pan, F. Metze, and M. Bett. Advances in Meeting Recognition. Submitted to HLT 2001, January 2001.

[12] M. Woszczyna, M. Broadhead, D. Gates, M. Gavaldà, A. Lavie, L. Levin, and A. Waibel. A Modular Approach to Spoken Language Translation for Large Domains. In *Proceedings of Conference of the Association for Machine Translation in the Americas (AMTA'98)*, Langhorn, PA, October 1998.

[13] T. Zeppenfeld, M. Finke, K. Ries, and A. Waibel. Recognition of Conversational Telephone Speech using the Janus Speech Engine. In *Proceedings of the ICASSP'97*, München, Germany, 1997.