9-1-1: Speech Translation for Emergency Dispatching

January 8, 2004

1 Introduction

We propose ...

This domain is very interesting, in that it is challenging but feasible. Challenging because ... Feasible because ... It also has clear social value. In addition to its specifically addressing the chronic need for Spanish translation at emergency dispatch centers (and in larger cities, a large number of other languages), this project is directly relevant to Homeland Defense. In any major natural or anthropogenic disaster, nearby 9-1-1 dispatching centers will receive large volumes of calls in a short period of time. In large cities, this will naturally include a significant number of non-English calls. If as we hope this research eventually leads to deployable speech translation systems for 9-1-1 dispatching, ...

2 The 9-1-1 Domain

A brief overview of the 9-1-1 dispatching task is in order here. Our interest in this potential domain for the application of speech translation technology began when Officer Julio Schrodel of the Cape Coral Police Department contacted us in October 2003, inquiring whether our previous speech translation work might help them with their translation needs. Since the availability of an enthusiastic user organization is an important asset in projects of this sort, we decided to investigate this possibility. We have had a number of follow-on discussions since then, and in December 2003 visited their police station and 9-1-1 dispatching center.

When someone dials **9-1-1** in Cape Coral, as in most places in the United States, they are connected to a special dispatching center. The operators there have been trained to perform a rapid *triage*, or categorization of emergency calls. As described to us (see next subsection), the initial decision is whether to send police, fire, or medical personnel. The appropriate units are dispatched as soon as this decision is made.

The dispatcher's task is not completed with this initial decision, however. While the responding unit is traveling to the location of the emergency, the dispatcher attempts to elicit more details about the emergency from the caller; they also try to keep the caller talking on the line, and to keep them calm. The additional details are intended to aid the responding unit, to help them prepare for the situation that they will encounter. As just one example, in the case of police responding to a call, the level of violence involved is very helpful to know in advance. Should they enter a building with guns drawn, or would that needlessly frighten or endanger someone simply reporting a missing child? Another important issue is the exact location of the emergency; although the 9-1-1 equipment automatically displays the phone number and address of the call, the information is not always correct, the emergency may not be at the location of the telephone, and cellular telephones do not (yet) provide 9-1-1 location information. As the dispatcher elicits more information about the emergency, they communicate it by radio to the responding unit.

9-1-1 dispatchers consciously follow a decision tree in this task[1]. Some example phrases that 9-1-1 dispatchers use are shown in Figure 1 (this is from the Cape Coral TDD (Telecommunications Device for the Deaf) unit).

[1] We are in the process of obtaining the full decision tree used in Cape Coral, beyond the sample shown in Figure 1.

```
911 what is your emergency
police fire or ambulance
what is your address
what is your telephone number
is this an apartment
we have help on the way
is it physical or verbal [abuse]
are there injuries
are there any weapons
cape coral police and fire department how may i help you
```

Figure 1: Sample 9-1-1 dispatcher questions (from Cape Coral PD)

2.1 The translation situation in Cape Coral

As in many other localities, the Cape Coral 9-1-1 center often receives emergency calls in languages other than English, primarily Spanish. When this happens during the daytime, the dispatching center connects the call to the Language Line human translation service[2] to translate for them. Obviously, many 9-1-1 calls arrive during the evening or night shift. At these times, a single on-duty bilingual police officer per shift handles the call center's Spanish translation needs, via 3-way calls with the officer.

While this arrangement is minimally adequate, it strains the department's resources, and they would be very interested in a semi-automated translation system. The Cape Coral Police Department intends to be an **Unfunded Collaborator** in this project, and has written a letter to that effect (see the **Supplemental Documents** section of this proposal). Cape Coral is a fairly affluent, medium sized town (about 160,000 residents), and seems to be a fairly ideal user partner in this enterprise. In addition to being enthusiastic about providing us with information about their operation, and eventually providing a live test site, they are willing to give us access to their recordings of actual 9-1-1 calls involving Spanish translation, for use as training data. While the main language issue in Cape Coral is Spanish (in many dialects, including European visitors), they also receive 9-1-1 calls in German and Canadian French. This would provide an opportunity in follow-on projects to test multi-lingual translation systems, although the current project will only work on Spanish-English translation.

2.2 Interesting characteristics of the 9-1-1 domain

The use of speech translation technology for the 9-1-1 domain is very interesting, in that it is definitely challenging, but apparently feasible. It is challenging due to:

- **Real-time requirements:** A deployed system would be in a real-time environment, where speed is of the essence, requiring rapid turn-around, both in the automated components and in the overall human dialogue loop[3].

[2] http://www.languageline.com/

[3] Of course, the current situation, connecting to a third-party translator, takes non-zero time, so some small delay would be acceptable.

- **Stressed speech:** The callers are often in a very emotionally distressed state, which affects the acoustic features of their speech, requiring new approaches to speech recognition.
- **Multiple dialects:** The Spanish callers will speak a variety of dialects, and may even include some English words and phrases.
- **Speech translation:** The translation of spoken language is still fundamentally difficult in any domain, with (to our knowledge) no actually deployed systems to date that perform "full" recognition and translation (as opposed to word-for-word or fixed-phrase systems).

Despite this perhaps daunting list of challenges, we believe that this domain is still actually feasible because:

- **Speech data availability:** 9-1-1 dispatchers routinely record all phonecalls, and keep the last 90 days on tape. The Cape Coral Police Department has indicated that they would have no problem providing us with these taped conversations for use as training data[4].
- **Strong task constraints:** As described above, the 9-1-1 dispatcher has a few specific, concrete goals in the interaction. These are to perform a top-level triage, followed by eliciting further specific (pre-defined) information depending on the nature of the emergency, as well as keeping the caller on the line and calm. This provides helpful constraints for the speech and translation components, as we explain below.
- **One-sided speech:** Another characteristic that simplifies this effort enough to be feasible is that the 9-1-1 dispatcher already is sitting at a computer workstation. The English side of the dialog can therefore be carried out in text, with English-to-Spanish information largely composed of whole pre-recorded phrases, most likely launched via menus or short commands. In addition to significantly reducing the amount of development required (perhaps by nearly 50%), a potential side benefit is a reduction in confusion (or cognitive load) by having only Spanish in a spoken modality, and only English in a written modality. Systems with two-way speech must grapple with issues (that we here avoid) of whether (and how) to let users hear the original other-language speech.
- **Our prior work:** We are fortunate here at CMU/LTI to have a great deal of experience in speech translation and multilingual speech work, much of it with Spanish, including a large technology base (both software and databases) and faculty and graduate students experienced in designing, implementing, and evaluating such systems. We will describe below how the new science in this proposal is based partly on prior NSF-funded research.

A final point to be made about the domain is the need to retain the human being in the dialogue loop. It seems clear to us that a naive person in a crisis situation would not be happy to deal with a fully automatic system; they will want to have a human being in the loop. Having the person in the loop allows us to make use of the human dispatcher's domain reasoning skills. We also envision (in eventual tests of the system) having a fall-back to the current human translator situation, in cases where the automatic system is not performing adequately. This system would still be highly valuable, in terms of reducing the load on the scarce human translator resources.

[4]We are in the process of obtaining sample speech data, and have asked them to begin stockpiling data longer-term, in the event that this proposal is approved.

3 Our 9-1-1 Speech Translation System Design

3.1 System architecture

Based on the 9-1-1 dispatching domain characteristics described in the preceding section, we are proposing a highly assymetrical overall system design.

In the **dispatcher-to-caller** direction, we will build a (relatively) simple phrase-based text-to-speech translation system. This will require no English automatic speech recognition (ASR), no (or very simple) true English-to-Spanish machine translation (MT), and relatively simple domain-limited Spanish Speech Synthesis. This should be even simpler (in terms of basic technology) than the Phraselator[30] described in section 6 below, although it still must be designed to work smoothly within the dispatcher's environment, and so will not be totally trivial.

The **caller-to-dispatcher** direction is much more interesting. In this direction we propose to develop:

- a novel Spanish speech recognizer that can handle emotional spontaneous speech of mixed dialects over the telephone, designed to interface to our translation system, and
- a novel Spanish-to-English translator that will be designed to make use of the dispatcher's decision tree and task constraints.
- (And no English speech synthesis at all.)

3.2 Project workplan

Good to include what happens in which years... [no quarterlies though]. Mention what students and UG will do.

- Acquire 9-1-1 decision tree, recordings of actual 9-1-1 calls
- Transcribe recordings, US English and Spanish (several dialects, three-way with interpreter).
- Decide specific $S->E$ MT approach based on transcripts, 9-1-1 tree. Probably either EBMT or SMT-variant; internal existing tech base for both Combined with simple limited-domain high-Q engine?? Classifier idea? [Similar to Knight's Babylon approach]
- Build initial pilot version of system
- Evaluate pilot version with CCPD
- Update system from test
- Evaluate alpha version, possibly with CCPD

There has been little evaluation in real situations (Tongues; Nespole; any others?)

"We don't foresee it being necessary, but we'll get IRB approval if it becomes appropriate."

Mention our experience in Tongues, Nespole! evaluations. Refer to Babylon future evaluation somewhere.

We'll evaluate system in real environment. But no live use yet: Test with real data off-line; Experts role-playing.

As mentioned in the preceding section, the Cape Coral Police Department is enthusiastic about this project, and hopes to eventually be the first test site in a follow-on project.

3.3 Broader Impacts

Actually, fits best at end of section 2??

[include this with social benefits: "In addition, the project will support two doctoral students, ... teaching, publishing"] – as will become clear from later sections, the funding requested in this proposal would advance the state of fundamental research in speech recognition and machine translation

Addressing Social Goals(??)

Should be an "integral part of the narrative"...

Social aspects: Chronic low-level need, plus Natural disasters and Homeland Security (low key) [already mentioned in domain description]

We will now describe in turn our approaches to the Speech Recognition, (very limited) Speech Synthesis, and Machine Translation component technologies in our proposed system.

4 Automatic Speech Recognition

As described above, the main goals in the 9-1-1 dispatching task are (1) to perform a rapid *triage*, or categorization of emergency calls, (2) to elicit more details about the emergency from the caller, and (3) to keep the caller on the line, and calm.

The nature of this task is highly demanding with respect to both the acoustic modeling and the language modeling of the speech recognition component of our proposed system. The challenges are: (1) real-time behaviour is required, (2) due to the nature of the 9-1-1 dispatcher task, we anticipate Spanish speech spoken over the telephone by highly distressed callers, (3) the spoken Spanish might be affected by a large variety of different dialects, and (4) since the callers live in the U.S., and as a consequence of the high distress of the speakers, we further expect the language may be partly intermixed with English words or phrases. (As descreibed above, only Spanish speech recognition is necessary, since the English-speaking dispatcher is already sitting at a computer console.)

In order to accomplish the triage task, to calm down the caller and keep him or her on the line, we believe that what is required is not a *full transcription* of the speech (describing in full, word-for-word detail what the caller was saying), but rather the main concepts of what the caller is intending to say. In other words, we propose to do here recognition using *grammar based concepts*. Decoding via a grammar improves the robustness of recognition, to counterbalance the challenging highly emotional speech, and also restricts the search space of the decoder, thereby increasing the decoding speed.

However, the given task also requires high flexibility, i.e. the decoder needs to handle ill-formed sentences, highly spontaneous speech, mix languages, and out-of-domain utterances. Simple grammar based recognition would not accomplish this; therefore, we propose to combine the robust grammar based recognition via context free grammars with the flexibility of statistical n-gram models. Furthermore, since the chances of language intermix between English and Spanish on the sentence level, and even phrase level, are high, we plan to employ multilingual grammars.

Decoding using a combination of statistical n-gram language models (LMs) and multilingual context based grammars will increase the robustness of the system; however, we also need special treatment for the acoustic models of the system, in order to handle speech spoken by highly distressed speakers from various Spanish dialects. We propose therefore to apply multilingual acoustic models enhanced by articulatory features. We expect those models to be more robust

towards dialectal variations and to emotional speech.

4.1 Janus Speech Recognition System

The JANUS speech recognition toolkit (JRTk), developed at the Interactive Systems laboratories (ISL) between 1993 and 2001[10, 33], recognizes speech in (often ill-formed) spontaneous, conversational spoken dialogues. Our efforts in large vocabulary (60,000+ words) speech recognition aim at greater robustness, rapid deployment and application to new domains and languages. Another emphasis is our effort to handle speech spoken by native and non-native speakers, as well as sloppy, conversational speech under noise and/or cross-talk, such as in the car, on the telephone or in conference rooms. The JANUS recognition toolkit was applied and ranked first in the official 96 and 97 DARPA Hub-5 benchmarks (conversational telephone speech) and all official German Verbmobil benchmark tests in 94 through 2000[41, 10, 40].

4.2 Multilingual Acoustic Models enhanced by Articulatory Features

Current Maximum Likelihood-trained speech recognition systems use *context-dependent acoustic models*, which supposedly capture the pronunciation variability by providing different pronunciation models for the same phone in different contexts, generated by a data-driven clustering technique. In this technique, a transcription of the speech is aligned with the speech waveform, with each part of the utterance assigned exactly one phone ("beads-on-a-string model"[26]), so that the corresponding model can be trained on these data. These models, however, will be very broad, because other factors, such as speaker accent, speech rate, age, gender, or noise conditions, also influence the acoustic representation of a sound. If, on the other hand, these factors are included in the clustering process, the training data is fragmented even more, so that every model is based on very little training data, and the resulting combinatorial explosion makes it difficult to compute useful shared models.

A richer description of the articulatory process than the phone sequence is given by it phonetically motivated features[8], which describe the articulatory state assumed by the speaker's vocal tract when uttering the sound, using features such as "VOICED" or "BILABIAL"/citeKirchhoff00. A feature vector then describes each phone, and speech recognition is based on recognition of the state sequence in the parallel feature streams. This approach allows the describtion of speech at a finer level than the phonetic segment, because not all features need to change synchronously between phones. Instead, a trajectory in multi-dimensional articulatory space describes the speech. Therefore *articulatory feature enhanced* speech recognition has much more potential to give valuable feedback on the pronunciation. Our initial experiments, in which we combined feature streams with conventional acoustic models trained on the Broadcast News corpus, showed that the error rate of the system using articulatory features decreased by up to 16%[25]. The same approach results in an error rate reduction of more than 25% in the evaluation of hyper-articulated data[34]. Our results on multilingual and crosslingual articulatory features gave up to 12.3% error rate reduction[36]. In addition, other groups have reported gains on small systems using feature-based approaches in noisy environments[9].

4.3 Speech recognition based on a combination of multilingual grammars and n-gram models

We propose a new language modeling method that allows us to combine several monolingual language models (LMs) into a single multilingual LM. Furthermore, in this work this technique will be extended to the concept of grammars. In order to allow seemless switching between languages, all linguistic knowledge sources share one common interface to the speech recognizer. We will leverage our work on multilingual acoustic models to develop truly multilingual speech recognition. This technique will enable us to recognize Spanish speech spoken with intermixed English phrases. The technological innovations here are that:

- n-gram LMs will be combined at the meta level without major loss in performance compared to monolingual ones,
- grammars will be applied to model multilinguality in limited domains,
- language switches will be significantly reduced,
- almost no resource overhead occurs for handling multiple languages in one language model, and
- language identification is done implicitly during decoding.

Using grammars instead of n-gram language models is especially an advantage in small domains, where less domain-dependent training data is available for n-gram language models. Rather than compiling one finite state graph out of all the terminals given by the grammars, we will use a more dynamical approach, where several rule-based finite state graphs, consisting of terminals and non-terminals, are linked together by their non-terminal symbols.

Several domain dependent sub-grammars can be activated/deactivated and loaded at run time. The grammars can also be expanded on the fly by new rules or terminals without restarting the recognizer. Even new words can be added to the grammar and the search network on the fly. These functionalities ensure that the recognizer can be rapidly adapted to the dispatcher level derived from the dispatcher's decision tree. To cope with spontaneous speech events, which occur frequently in distressed speech, we will be using the mechanism of filler words in the decoder, which can potentially occur between any two terminals. Instead of asking the language model for their score, a predefined filler penalty is applied.

We propose to combine this robust grammar-based recognition via context free grammars with the flexibility of statistical n-gram models. This will be realized by embedding grammar-based phrase decoding into n-gram based language models.

In recent experiments [11] we recognized utterances from multiple languages with the use of a single recognition engine. This not only requires a multilingual language model but also a multilingual acoustic model; the latter problem has already been extensively studied [31] and the results applied to these experiments. Our results proved that multilingual grammars can be used to efficiently decode two languages (English and German) within a single system. The results also showed that n-gram language models can be combined at a meta-level and thus allow the preservation of language-specific information captured in the individual language models. The resulting system is easier to maintain, allows decoupled optimization, and implicitly identifies the language spoken during decoding. Moreover, we showed that the grammar based system is roughly twice as fast as the n-gram LM, proving that decoding along context free grammars in such restricted domains

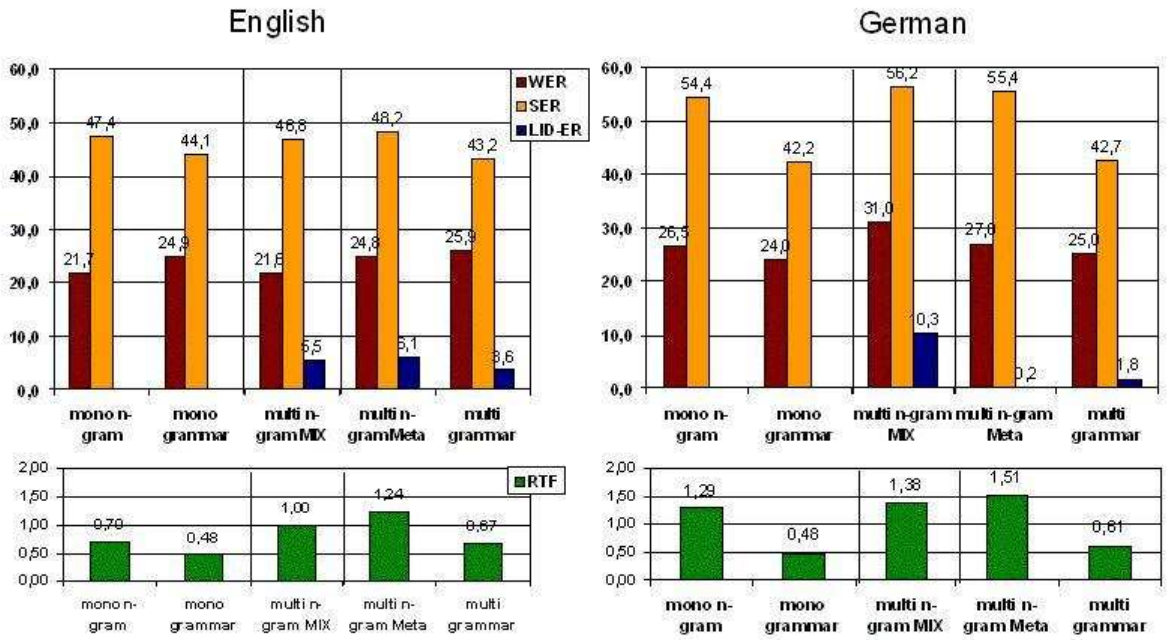

Figure 2: Comparison of ASR Performance for various Types of Multilingual Grammars

gives a large advantage over the standard n-gram approach (see Figure 2). In sum, the combination of multilingual language models and multilingual acoustic model allows seamless switching between languages (and domains).

5 Speech Synthesis

Speech translation systems require both recognition and synthesis. Within this project, we do not aim to include any new fundamental research in speech synthesis, but we still require some control over the spoken Spanish output. At CMU/LTI we are lucky to be one of the world centers in speech synthesis, and we will leverage tools and techniques from our FestVox synthesis group in this project.

For the most part, textual output in English is likely to be sufficient for communication, as the English-speaking dispatcher will be familiar with the system. Also, in some cases only partial translations may be produced, so text may well be the most reasonable form of English output.

For the Spanish output, the desired content will often be fixed, or nearly-fixed, format. In such situations, if simple pre-recorded Spanish output is not suitable, because there is too much variation or the output utterances are not yet well enough defined, the best solution is *limited domain synthesis* [4]. In limited domain synthesis, the set of utterances that may be spoken is well-defined (though perhaps vary large or even infinite). Given the set of utterances, well-defined procedures and scripts are available to find the best subset of training sentences that covers the phonetic and prosodic space. These can be recorded, and a high quality voice can be built with relatively little effort [3]. Such a limited domain system for Spanish is only feasible for us due to the

tools already developed here and the existence of a full, general, Latin-American Spanish synthesis voice to bootstrap from.

There are other issues in voice output that may become relevant at later stages of this research, in subsequent follow-on projects. For example, style of output may be important. In stressed conditions, it may be important that the output voice is a command, e.g., "Leave the building now", or compassionate, "An ambulance is on its way". Such stylistic variation can be covered within a limited domain unit selection framework [2], but only when it is constrained, as it should be in this project.

6 MT Research

The language translation tasks that are required in the 9-1-1 domain, as have been outlined above, are unique, and assymetrical:

- In one direction, we need to provide translation of Spanish utterances spoken by an untrained 9-1-1 caller into English, but only well enough so that an English-speaking 9-1-1 dispatcher can decide what sort of emergency is being reported (at several levels of classification in the dispatcher's decision tree), and can then gather crucial relevant details, such as an address.
- In the other direction, relatively fixed English inquiries and requests from the trained 9-1-1 dispatcher need to be conveyed in Spanish back to the caller.

Each of these two directions is quite different from a "standard" Machine Translation (MT) task, which would normally require full translation of freely-varying text from the source language to the target language. Our approach to these two directions therefore will be quite different:

- In the **dispatcher-to-caller** direction, we will primarily rely on lookup-based translation of fixed-phrase expressions, similar to the "Phraselator"[30] that has been developed under DARPA auspices. The Phraselator system leverages strong assumptions about its use: that the operator is trained to only use specific phrases, and that the (untrained) hearer's speech does not need to be translated. By recognizing whole source language phrases, and playing whole (pre-recorded) target language phrases, the quality of the system's output (within this narrow application) can be very high, and the system can be put together with relatively miniscule effort. This scenario is actually a good fit to the dispatcher-to-caller direction in the 9-1-1 domain.
- In the **caller-to-dispatcher** direction, our goal is to provide MT-supported real-time translation from unconstrained Spanish into English. As already noted, the challenges in this direction are significant: the input for translation is Spanish text provided by a speech recognizer, from telephone speech from a distressed caller. By the nature of the task, we expect this input to be extremely noisy and ungrammatical. The task also, however, has several significant constraints, which we can use to our advantage: the domain is highly limited, and we do not require a full translation. This is where the meat of our MT research activities will be.

But first, a word from our sponsor:

The PIs of this proposed project have broad research experience in the development of MT systems from a wide range of research projects on text and speech translation over the past decade. Interestingly, the technology we propose to develop in this project stems to a significant degree from novel extensions of two prior NSF-funded projects, which we describe in the next subsection.

6.1 Results from Prior NSF-Supported Work

NESPOLE![5] (NSF-9982227, 03/01/00–02/28/03, $1,499,217) is a speech-to-speech machine translation project funded jointly by the European Commission and the US NSF under the MLIAM program. The main goal of NESPOLE! is to advance the state-of-the-art of speech-to-speech translation in real-world settings. The main research challenges are improving system robustness in practical environments and developing new translation methods that are easier to port to new and broad domains. The system is designed to provide fully functional speech-to-speech capabilities within real-world settings of common users involved in e-commerce applications. The Interlingua-based translation system covers the domain of tourist information and travel planning, and the domain of medical assistance. The main results of the project include a distributed architecture for speech translation that allows flexible physical distribution of modules over the internet; an integrated multi-modal interface that supports communication via shared web pages, whiteboard images and video-conferencing in which speech-translation is seamlessly embedded; a new interlingua representation framework that is reliable for multi-lingual development and is easier to expand into new domains; and new data-trainable analysis approaches that reduce the amount of expert human labor required for system development. Showcases of the system have been demonstrated at HLT-2002, AMTA-2002 and IST-2002 in Berlin. Resulting publications include [37], [17], [14], [22], [18], [7], [15], [16],[5], [21]. The project is a collaboration between three European research laboratories (IRST in Trento, Italy; ISL at Universität Karlsruhe (TH) in Germany; and CLIPS at Université Joseph Fourier in Grenoble, France), one US research group (ISL at Carnegie Mellon University in Pittsburgh, PA) and two industrial partners (APT; Trento, Italy – the Trentino provincial tourism board, and Aethra; Ancona, Italy – a tele-communications company).

The **AVENUE** Project (NSF-0121631, 09/01/01–08/31/06, $2,500,001) is a five-year project currently in its third year, funded by the ITR initiative. Its primary research goal is to develop speech and translation methods for native languages spoken by minority populations, for which very limited amounts of data and resources are available. The main new translation approach is a method for acquiring high-quality MT transfer rules from native informants, which decreases dependence on human experts and reduces development time. Semantically-conditioned transfer rules are generalized via a new locally-constrained Seeded Version-Space method based on a controlled bilingual corpus and interactive tools to elicit information from native informants. New speech recognition research builds general phone models across multiple language families and adapts the recognizer to new languages with minimal new-language training data. All of these methods are based on new and existing machine learning algorithms that combine prior knowledge with limited amounts of new data in order to converge quickly on working machine translation and speech recognition systems. The project has been actively working on Mapudungun, a native language of southern Chile (a collaboration with a research team in Temuco, Chile), and is preparing to expand to Quechua (spoken in Ecuador and Peru), and Inupiaq (a native language of Alaska). Resulting publications include [6], [20], [27].

6.2 Our proposed new Machine Translation research

Since we have a well-defined domain, in terms of states in the dispatcher's decision tree, we plan to apply our previous work on *Domain Action* classification[23, 13]. We use the term "Domain Action" (DA) to refer to the combination of a general Speech Act with domain-specific concepts. Domain Actions capture *speaker intention* in a limited-domain system, rather than detailed *literal meaning*,

[5] NESPOLE! – NEgotiation through SPOken Language in E-commerce.

and have been used successfully in the NESPOLE! and C-STAR speech translation projects. We will adapt and extend this prior work in order to classify each recognized utterance into a DA. One extension we plan to investigate is to condition the DA classification of the *caller's* utterance based on the *dispatcher's* state in the decision tree, to capture the expectation that, e.g., if the dispatcher has asked for an address, the caller is likely to provide one (or produce another relevant Domain Action, such as "I don't know!")

Once we have classified the caller's utterance, the goal of the next translation stage is to provide, in English, more detailed information beyond the initial classification, although not typically a full literal translation of the original Spanish utterance. In particular, we will be trying to identify and translate salient pieces of information that are pertinent to a specific DA in the 9-1-1 domain. We refer to these DA-specific pieces of information as semantic "arguments" (as in the arguments of a verb or function).

The translation of these arguments will be implemented by adapting and extending an existing well-developed transfer-based framework that was created at the LTI under the NSF-funded AVENUE Project [6], [20], [27]. Our transfer-rule formalism supports succinct representation of the necessary information for the transfer engine to perform analysis, transfer and generation at run-time. The main elements of one of our transfer rules include:

- a set of source and target **transfer components**,
- **alignments** between source and target components, and
- a set of **unification constraints** that can be used to constrain the applicability of the rule, and to transfer feature information from source to target language.

The transfer components in a rule can be lexical items or higher-level constituents, the transfer of which is supported compositionally by lower-level rules.

Our transfer rule formalism is able to handle a variety of common translation divergences, including:

- **head-switching**, where a different word is the head of the corresponding phrase in the two different languages,
- **thematic changes**, such as an object in one language being expressed as a subject in the target language,
- **structural changes**, such as having an NP become a PP in another language, and
- **lexical gaps**, where one target word replaces an entire source phrase [38].

Some examples of this type of transfer rule for Chinese-English are shown in Figure 3.

The main steps in the development of this transfer grammar are the following:

1. We first collect a corpus of Spanish utterances, hand-classify them into DAs, and identify the patterns of pertinent information that are characteristic of *each* DA. The information patterns are manually translated from Spanish to English, and word-aligned.
2. Once the arguments have been identified, tagged, and aligned, a student trained as a bilingual grammar developer will design transfer rules, specific to the DAs, that capture the specific patterns of arguments in both the source and target language, and how they correspond[6].

[6]The AVENUE project is developing learning techniques to automate this step, but they have not yet reached the point of being applicable to a project such as this one.

```
; Rule to handle non-auxiliary verb
; question transfer from Chinese to
; English
{S,3}
S::S : [NP VP MA] -> [V NP VP
"?"]
(;; Constituent alignments
(X1::Y2)
(X2::Y3)
;; x-side constraints ;; (parsing)
((x0 subj) == x1)
((x0 subj case) = nom)
(x0 = x2)
((x0 act) = quest)
;; xy-constraints ;; (transfer)
(y0 = x0)
((y0 act) =c quest)
;; y-side constraints ;; (generation)
((y1 form) = do)
((y1 agr) = (y2 agr))
(y2 == (y0 subj))
(y3 = y0))
```

```
; Sample lexical transfer rules
AUX::AUX |: [zuo4] -> [do]
((x1::y1)
((y0 form) = do)
((y0 agr) = (*or* 1sg 2sg pl))
((y0 tense) = present)
)

AUX::AUX |: [zuo4] -> [does]
(
(x1::y1)
((y0 form) = do)
((y0 agr) = 3sg)
((y0 tense) = present))
```

Figure 3: Sample Transfer Rules

3. At run-time, a partial parser/matcher will be used in order to identify matches of source-language arguments that are specified in the transfer grammar. The transfer engine will then support the analysis, transfer and generation into English of the matched arguments, according to the transfer rules. The main issues involved in developing the partial parser/matcher are efficiency of search and effective handling of high-levels of ambiguity. We have a great deal of experience in building this style of parser for speech applications[19, 29], as well as having a significant pre-existing code base available to be leveraged.

Note that, due to the design of the Automatic Speech Recognition (ASR) in this system (see section 4), the output of ASR may consist of a mix of parsed fragments and surface strings. This is actually fairly unproblematic, since this resembles the state of a partially-parsed sentence that would occur during normal parsing. It simply obviates the need to parse the segment which ASR has pre-parsed.

A common criticism of rule-based translation systems is that they require an inordinate amount of human grammar-writing effort, making them prohibitively expensive. We do not expect this to be a major problem in this project, because for this domain we will only be parsing a small set of arguments, and will have strong, state-dependent expectations as to what will be parsed in any given utterance (e.g., an address, a spelled name, a type of injury). We expect that a number of small, non-interacting, state-specific argument grammars will not be prohibitively difficult to produce.

7 Other Related Work

We have already discussed above the DARPA-sponsored Phraselator[30] system, and our own prior NSF-sponsored related work. We will discuss in this section other related work.

The most closely related other current work is probably the DARPA-supported work under the Babylon program. This includes a two-way Arabic-English system produced here called the Speechalator[39]. The Speechalator differs in a number of ways from what we are proposing here, due to a fairly different target environment. It is intended for face-to-face dialogues in a doctor-patient domain, and is designed to run on a commercial PDA. No special considerations were made for dealing with emotional speech or mixed dialects, and there was an implicit assumption that the utterances on both sides would remain within the covered domain (while in this proposal, we assume we can only translate a (critical) portion of what is said).

Other work in the BABYLON program shares some surface similarities to our proposed work. The BBN Babylon system[24], apparently based on earlier BBN HMM-based work[32], is designed to do a sort of extraction and translation, but based on statistical training, and not using a domain-based classification step. While statistical training has shown great success in text-based applications, we do not believe it is promising for this sort of speech application.

The other BABYLON work with some apparent similarity is that of the Hughes/ISI group[12]. In this work, they do perform domain-class-based classification, but they do not perform any spot translation of critical information. This system is again entirely statistically trained.

The other class of related work might be from the automated call center area of research. (Although an important feature of our project is that we want to translate for a human dispatcher who is still in the dialogue loop.) There has been work on monolingual, fully-automated call centers for some time now[28]. The one such system that we are aware of that attempts translation, Anuvaad[1], again uses statistical techniques, and resorts to attempting to train from the output of commercial MT systems. We feel this is an unlikely path to success, since it seems such a system will learn to perform as badly as general-purpose MT systems. We are aware of one other system in this class that bears mentioning: the AMITIES[35] effort to construct automated multilingual call center applications. Note that a multilingual call center is actually different from a translation-enabled call center; also this project seems to be somewhat stalled due to the difficulty of acquiring sufficient domain speech data.

References

[1] Srinivas Bangalore and Giuseppe Riccardi. A Finite-State Approach to Machine Translation. In *Proceedings of the North American ACL 2001 (NAACL-2001)*, Pittsburgh, May 2001.

[2] Alan Black. Unit selection and emotional speech. In *Eurospeech*, Geneva, Switzerland., 2003.

[3] Alan Black and Kevin Lenzo. Building voices in the Festival speech synthesis system. http://festvox.org/bsv/, 2000.

[4] Alan Black and Kevin Lenzo. Limited domain synthesis. In *ICSLP2000*, volume II, pages 411–414, Beijing, China., 2000.

[5] S. Burger, L. Besacier, P. Coletti, F. Metze, and C. Morel. The NESPOLE! VoIP Dialogue Database. In *Proceedings of EuroSpeech 2001*, Aalborg, Denmark, September 2001. ISCA.

[6] Jaime Carbonell, Katharina Probst, Erik Peterson, Christian Monson, Alon Lavie, Ralf Brown, and Lori Levin. Automatic Rule Learning for Resource Limited MT. In *Proceedings of 5th Conference of the Association for Machine Translation in the Americas (AMTA-2002)*, pages 1–10, Tiburon, CA, October 2002. Springer.

[7] Roldano Cattoni, Marcello Federico, and Alon Lavie. Robust Analysis of Spoken Input combining Statistical and Knowledge-based Information Sources. In *Proceedings of the 2001 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2001)*, Madonna di Campiglio, Italy, December 2001.

[8] Li Deng. A Dynamic, feature-based Approach to the Interface between Phonology and Phonetics for Speech Modeling and Recognition. In *Speech Communication, Vol 24, No.4*, pages 299–323, 1998.

[9] Eric Eide. Distinctive Features For Use in an Automatic Speech Recognition System. In *Proceedings of the Eurospeech, Aalborg, Denmark*, 2001.

[10] Michael Finke, Petra Geutner, Hermann Hild, Thomas Kemp, Klaus Ries, and Martin Westphal. The Karlsruhe-VERBMOBIL Speech Recognition Engine. In *Proceedings of the ICASSP*, Munich, 1997.

[11] Christian Fügen, Sebastian Stüker, Florian Metze, Hagen Soltau, and Tanja Schultz. Efficient Handling of Multilingual Language Models. In *Proceedings of ASRU*, 2003.

[12] Kevin Knight. HRL Technical Briefing. Presentation at DARPA BABYLON PI Meeting, December 2003.

[13] Chad Langley. *Domain Action Classification and Argument Parsing for Interlingua-Based Spoken Language Translation.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2003.

[14] Chad Langley, Alon Lavie, Lori Levin, Dorcas Wallace, Donna Gates, and Kay Peterson. Spoken Language Parsing using Phrase-level Grammars and Trainable Classifiers. In *Proceedings of Speech-to-Speech Translation Workshop at the 40th Annual Meeting of the Association of Computational Linguistics (ACL-02)*, pages 15–22, Philadelphia, PA, July 2002.

[15] Alon Lavie, Franco Balducci, Paolo Coletti, Chad Langley, Gianni Lazzari, Fabio Pianesi, Loredana Taddei, and Alex Waibel. Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-Commerce Applications. In *Proceedings of the 2001 Human Language Technology Conference (HLT-2001)*, pages 31–34, San Diego, CA, March 2001. DARPA.

[16] Alon Lavie, Lori Levin, Tanja Schultz, Chad Langley, Benjamin Han, Alicia Tribble, Donna Gates, Dorcas Wallace, and Kay Peterson. Domain Portability in Speech-to-Speech Translation. In *Proceedings of the 2001 Human Language Technology Conference (HLT-2001)*, pages 82–86, San Diego, CA, March 2001. DARPA.

[17] Alon Lavie, Florian Metze, Roldano Cattoni, and Erica Costantini. A Multi-Perspective Evaluation of the NESPOLE! Speech-to-Speech Translation System. In *Proceedings of Speech-to-Speech Translation Workshop at the 40th Annual Meeting of the Association of Computational Linguistics (ACL-02)*, pages 121–128, Philadelphia, PA, July 2002.

[18] Alon Lavie, Florian Metze, Fabio Pianesi, Sussane Burger, Donna Gates, Lori Levin, Chad Langley, Kay Peterson, Tanja Schultz, Alex Waibel, Dorcas Wallace, John McDonough, Hagen Soltau, Roldano Cattoni, Gianni Lazzari, Nadia Mana, Emanuele Pianta, Erica Costantini, Laurent Besacier, Herve Blanchon, Dominique Vaufreydaz, and Loredana Taddei. Enhancing the Usability and Performance of NESPOLE! - a Real-world Speech-to-Speech Translation System. In *Proceedings of the 2002 Human Language Technology Conference (HLT-2002)*, pages 269–274, San Diego, CA, March 2002. DARPA.

[19] Alon Lavie and Masaru Tomita. GLR* - An Efficient Noise-Skipping Parsing Algorithm for Context-Free Grammars. In H. Bunt and M. Tomita (eds.), editors, *Recent Advances in Parsing Technology*. Text Speech and Language Technology series (vol. 1), Kluwer Academic Press, August 1996.

[20] L. Levin, R. Vega, J. Carbonell, R. Brown, A. Lavie, E. Canulef, and C. Huenchullan. Data Collection and Language Technologies for Mapudungun. In *Proceedings of International Workshop on Resources and Tools in Field Linguistics at the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Canary Islands, Spain, June 2002.

[21] Lori Levin, Donna Gates, Fabio Pianesi, Donna Wallace, Takeshi Watanabe, and Monika Woszczyna. Evaluation of a Practical Interlingua for Task-Oriented Dialogues. In *Proceedings of Workshop On Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP, at ANLP/NAACL-2000 Conference*, pages 18–23, Seattle, WA, April 2000.

[22] Lori Levin, Donna Gates, Dorcas Wallace, Kay Peterson, Alon Lavie, Fabio Pianesi, Emanuele Pianta, Roldano Cattoni, and Nafia Mana. Balancing Expressiveness and Simplicity in an Interlingua for Task-based Dialogue. In *Proceedings of Speech-to-Speech Translation Workshop at the 40th Annual Meeting of the Association of Computational Linguistics (ACL-02)*, pages 53–60, Philadelphia, PA, July 2002.

[23] Lori Levin, Chad Langley, Alon Lavie, Donna Gates, Dorcas Wallace, and Kay Peterson. Domain Specific Speech Acts for Spoken Language Translation. In *Proceedings of 4th SIGdial Workshop on Discourse and Dialogue (SIGDIAL-2003)*, Sapporo, Japan, July 2003.

[24] John Makhoul. BBN Technical Briefing. Presentation at DARPA BABYLON PI Meeting, December 2003.

[25] Florian Metze and Alex Waibel. A flexible Stream Architecture for ACR Using Articulatory Feature. In *Proceedings of ICSLP, Denver, Colorado*, 2002.

[26] Mari Ostendorf. Moving Beyond The Beads-on-a-string Model Of Speech. In *Proceedings of ASRU*, 1999.

[27] Katharina Probst, Ralf Brown, Jaime Carbonell, Alon Lavie, Lori Levin, and Erik Peterson. Design and Implementation of Controlled Elicitation for Machine Translation of Low-density Languages. In *Proceedings of the MT-2010 Workshop at MT-Summit VIII*, Santiago de Compostela, Spain, September 2001.

[28] G. Riccardi, A.L. Gorin, A. Ljolje, and M. Riley. Spoken Language Understanding for Automated Call Routing. In *Proceedings of ICASSP-97*, Munich, Germany, April 1997.

[29] Carolyn P. Rose and Alon Lavie. Balancing Robustness and Efficiency in Unification-augmented Context-Free Parsers for Large Practical Applications. In van Noord and Junqua (eds.), editors, *Robustness in Language and Speech Technology*. ELSNET series, Kluwer Academic Press, 2001.

[30] Sarich, Ace. Phraselator, one-way speech translation system. http://www.sarich.com/translator/, 2001.

[31] Tanja Schultz and Alex Waibel. Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition. *Speech Communication*, 35, August 2001.

[32] Richard Schwartz, Scott Miller, David Stallard, and John Makhoul. Language understanding using hidden understanding models. In *Proc. ICSLP '96*, volume 2, pages 997–1000, Philadelphia, PA, 1996.

[33] Hagen Soltau, Florian Metze, Christian Fügen, and Alex Waibel. A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment. In *Proceedings of the ASRU*, Madonna di Campiglio Trento, Italy, December 2001.

[34] Hagen Soltau, Florian Metze, and Alex Waibel. Compensating for Hyperarticulation by modeling articulatory properties. In *Proceedings of ICSLP, Denver, Colorado*, 2002.

[35] Tomek Strzalkowski. AMITIES Special Presentation. Presentation at DARPA BABYLON PI Meeting, December 2003. http://www.dcs.shef.ac.uk/nlp/amities/.

[36] Sebastian Stüker, Tanja Schultz, Florian Metze, and Alex Waibel. Multilingual Articulatory Features. In *Proceedings of the ICASSP, Hong Kong, China*, 2003.

[37] Loredana Taddei, Erica Costantini, and Alon Lavie. The NESPOLE! Multimodal Interface for Cross-lingual Communication: Experience and Lessons Learned. In *Proceedings of IEEE International Conference on Multimodal Interfaces (ICMI-2002)*, Pittsburgh, PA, October 2002. IEEE.

[38] Arturo Trujillo. *Translation Engines: Techniques for Machine Translation*. Springer-Verlag London Limited, London, 1999.

[39] Alex Waibel, Ahmed Badran, Alan W Black, Robert Frederking, Donna Gates, Alon Lavie, Lori Levin, Kevin Lenzo, Laura Mayfield Tomokiyo, Juergen Reichert, Tanja Schultz, Dorcas Wallace, Monika Woszczyna, and Jing Zhang. Speechalator: two-way speech-to-speech translation on a consumer PDA. In *Proceedings of Eurospeech-2003*, Geneva, Switzerland, 2003.

[40] Alex Waibel, Hagen Soltau, Tanja Schultz, Thomas Schaaf, and Florian Metze. Multilingual Speech Recognition. *Chapter in: Verbmobil - Foundations of Speech-to-Speech Translation, Wolfgang Wahlster (Ed.)*, 2000.

[41] Torsten Zeppenfeld, Michael Finke, Klaus Ries, Martin Westphal, and Alex Waibel. Recognition of conversational telephone speech using the janus speech engine. In *Proceedings of the ICASSP, Munich, Germany*, 1997.