Speech Translation for Triage of Emergency Phonecalls in Minority Languages

Udhyakumar Nallasamy, Alan W Black, Tanja Schultz, Robert Frederking
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
udhay@cmu.edu,
{awb,ref,tanja}@cs.cmu.edu

Jerry Weltman
Louisiana State University
Baton Rouge,
Louisiana 70802 USA
jweltm2@lsu.edu

Abstract

We describe Ayudame, a system designed to recognize and translate Spanish emergency calls for better dispatching. We analyze the research challenges in adapting speech translation technology to 9-1-1 domain. We report our initial research in 9-1-1 translation system design, ASR experiments, and utterance classification for translation.

1 Introduction

In the development of real-world-applicable language technologies, it is good to find an application with a significant need, and with a complexity that appears to be within the capabilities of current existing technology. Based on our experience in building speech-to-speech translation, we believe that some important potential uses of the technology do not require a full, complete speech-to-speech translation system; something much more lightweight can be sufficient to aid the end users (Gao et al, 2006).

A particular task of this kind is dealing with emergency call dispatch for police, ambulance, fire and other emergency services (in the US the emergency number is 9-1-1). A dispatcher must answer a large variety of calls and, due to the multilingual nature of American society, they may receive non-English calls and be unable to service them due to lack of knowledge of the caller language.

As a part of a pilot study into the feasibility of dealing with non-English calls by a mono-lingual English-speaking dispatcher, we have designed a translation system that will aid the dispatcher in communicating without understanding the caller's language.

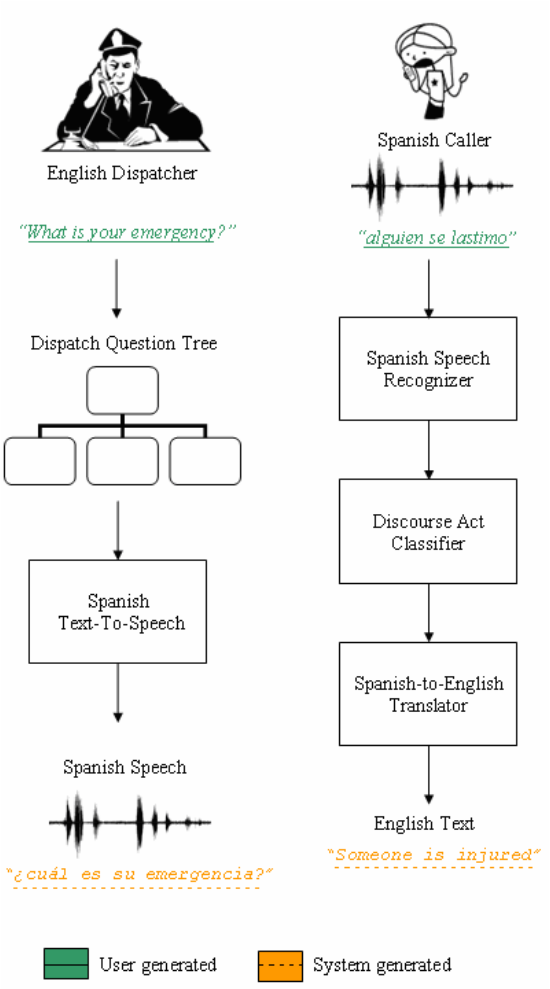

Figure 1. Ayudame system architecture

The fundamental idea is to use utterance classification of the non-English input. The non-English is first recognized by a speech recognition system; then the output is classified into a small number of domain-specific classes called Domain Acts (DAs) that can indicate directly to the dispatcher the general intended meaning of

the spoken phrase. Each DA may have a few important parameters to be translated, such as street addresses (Levin et al, 2003; Langley 2003). The dispatcher can then select from a limited number of canned responses to this through a simple menu system. We believe the reduction in complexity of such a system compared to a full speech-to-speech translation will be advantageous because it should be much cheaper to construct, easier to port to new languages, and, importantly, sufficient to do the job of processing emergency calls.

In the "NineOneOne" project, we have designed an initial prototype system, which we call "Ayudame" (Spanish word for "Help me"). Figure 1 gives an overview of the system architecture.

2 The NineOneOne Domain

Our initial interest in this domain was due to contact from the Cape Coral Police Department (CCPD) in Florida. They were interested in investigating how speech-to-speech translations could be used in emergency 9-1-1 dispatch systems. Most current emergency dispatching centers use some proprietary human translation service, such as Language Line[1]. Although this service provides human translation services for some 180 languages, it is far from ideal. Once the dispatcher notes that the caller cannot speak/understand English, they must initiate the call to Language Line, including identifying themselves to the Language Line operator, before the call can actually continue. This delay can be up to a minute, which is not ideal in an emergency situation.

After consulting with CCPD, and collecting a number of example calls, it was clear that full speech-to-speech translation was not necessary and that a limited form of translation through utterance classification (Lavie et al, 2001) might be sufficient to provide a rapid response to non-English calls. The language for our study is Spanish. Cape Coral is on the Gulf Coast of Florida and has fewer Spanish speakers than e.g. the Miami area, but still sufficient that a number of calls are made to their emergency service in Spanish, yet many of their operators are not sufficiently fluent in Spanish to deal with the calls.

There are a number of key pieces of information that a dispatcher tries to collect before passing on the information to the appropriate emergency service. This includes things like location, type of emergency, urgency, if anyone is hurt, if the situation is dangerous, etc. In fact many dispaching organizations have existing, well-defined policies on what information they should collect for different types of emergencies.

3 Initial system design

Based on the domain's characteristics, in addition to avoiding full-blown translation, we are following a highly asymmetrical design for the system (Frederking et al, 2000). The dispatcher is already seated at a workstation, and we intend to keep them "in the loop", for both technical and social reasons. So in the dispatcher-to-caller direction, we can work with text and menus, simplifying the technology and avoiding some cognitive complexity for the operator. So in the dispatcher-to-caller direction we require

- ***no*** English ASR,
- ***no*** true English-to-Spanish MT, and
- simple, domain-limited, Spanish speech synthesis.

The caller-to-dispatcher direction is much more interesting. In this direction we require

- Spanish ASR that can handle emotional spontaneous telephone speech in mixed dialects,
- Spanish-to-English MT, but
- ***no*** English Speech Synthesis.

We have begun to consider the user interfaces for Ayudame as well. For ease of integration with pre-existing dispatcher workstations, we have chosen to use a web-based graphical interface. For initial testing of the prototype, we plan to run in "shadow" mode, in parallel with live dispatching using the traditional approach. Thus Ayudame will have a listen-only connection to the telephone line, and will run a web server to interact with the dispatcher. Figure 2 shows an initial design of the web-based interface. There are sections for a transcript, the current caller utterance, the current dispatcher response choices, and a button to transfer the interaction to a human translator as a fall-back option. For each utterance, the DA classification is displayed in addition to the actual utterance (in case the dispatcher knows some Spanish).

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

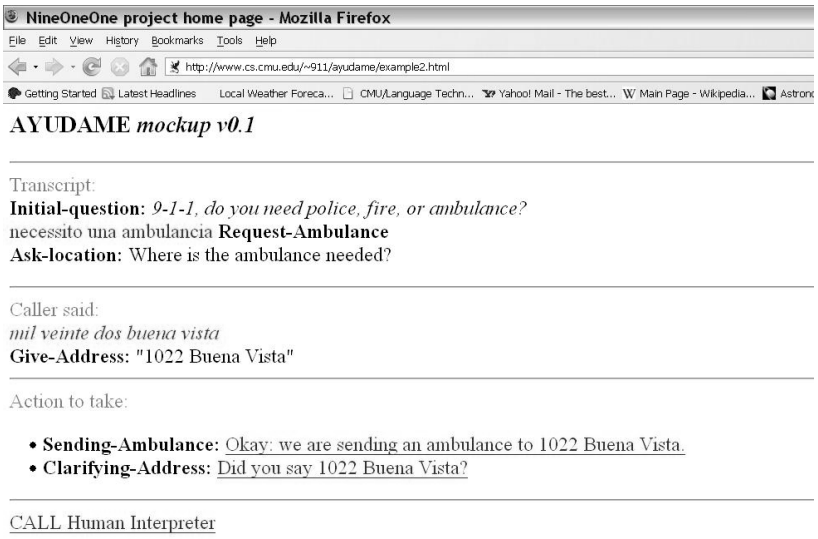

Figure 2. Example of initial GUI design

4 Automatic Speech Recognition

An important requirement for such a system is the ability to be able to recognize the incoming non-English speech with a word error rate sufficiently low for utterance classification and parameter translation to be possible. The issues in speech recognition for this particular domain include: telephone speech (which is through a limited bandwidth channel); background noise (the calls are often from outside or in noisy places); various dialects of Spanish, and potential stressed speech. Although initially we expected a substantial issue with recognizing stressed speakers, as one might expect in emergency situations, in the calls we have collected so far, although it is not a negligible issue, it is far less important that we first expected.

The Spanish ASR system is built using the Janus Recognition Toolkit (JRTk) (Finke et al, 1997) featuring the HMM-based IBIS decoder (Soltau et al, 2001). Our speech corpus consists of 75 transcribed 9-1-1 calls, with average call duration of 6.73 minutes (min: 2.31 minutes, max: 13.47 minutes). The average duration of Spanish speech (between interpreter and caller) amounts to 4.8 minutes per call. Each call has anywhere from 46 to 182 speaker turns with an average of 113 speaker turns per call. The turns that have significant overlap between speakers are omitted from the training and test set. The acoustic models are trained on 50 Spanish 9-1-1 calls, which amount to 4 hours of speech data.

The system uses three-state, left-to-right, subphonetically tied acoustic models with 400 context-dependent distributions with the same number of codebooks. Each codebook has 32 gaussians per state. The front-end feature extraction uses standard 39 dimensional Mel-scale cepstral coefficients and applies Linear Discriminant Analysis (LDA) calculated from the training data. The acoustic models are seeded with initial alignments from GlobalPhone Spanish acoustic models trained on 20 hours of speech recorded from native Spanish speakers (Schultz et al, 1997). The vocabulary size is 65K words. The language model consists of a trigram model trained on the manual transcriptions of 40 calls and interpolated with a background model trained on GlobalPhone Spanish text data consisting of 1.5 million words (Schultz et al, 1997). The interpolation weights are determined using the transcriptions of 10 calls (development set). The test data consists of 15 telephone calls from different speakers, which amounts to a total of 1 hour. Both development and test set calls consisted of manually segmented and transcribed speaker turns that do not have a significant overlap with other speakers. The perplexity of the test set according to the language model is 96.7.

The accuracy of the Spanish ASR on the test set is 76.5%. This is a good result for spontaneous telephone-quality speech by multiple unknown speakers, and compares favourably to the ASR accuracy of other spoken dialog systems. We had initially planned to investigate novel ASR techniques designed for stressed speech and multiple dialects, but to our surprise these do not

seem to be required for this application. Note that critical information such as addresses will be synthesized back to the caller for confirmation in the full system. So, for the time-being we will concentrate on the accuracy of the DA classification until we can show that improving ASR accuracy would significantly help.

5 Utterance Classification

As mentioned above, the translation approach we are using is based on utterance classification. The Spanish to English translation in the Ayudame system is a two-step process. The ASR hypothesis is first classified into domain-specific Domain Acts (DA). Each DA has a predetermined set of parameters. These parameters are identified and translated using a rule-based framework. For this approach to be accomplished with reasonable effort levels, the total number of types of parameters and their complexity must be fairly limited in the domain, such as addresses and injury types. This section explains our DA tagset and classification experiments.

5.1 Initial classification and results

The initial evaluation (Nallasamy et al, 2008) included a total of 845 manually labeled turns in our 9-1-1 corpus. We used a set of 10 tags to annotate the dialog turns. The distribution of the tags are listed below

Tag (Representation)	Frequency
Giving Name	80
Giving Address	118
Giving Phone number	29
Requesting Ambulance	8
Requesting Fire Service	11
Requesting Police	24
Reporting Injury/Urgency	61
Yes	119
No	24
Others	371

Table 1. Distribution of first-pass tags in the corpus.

We extracted bag-of-word features and trained a Support Vector Machine (SVM) classifier (Burges, 1998) using the above dataset. A 10-fold stratified cross-validation has produced an average accuracy of 60.12%. The accuracies of individual tags are listed below.

Tag	Accuracy (%)
Giving Name	57.50
Giving Address	38.98
Giving Phone number	48.28
Req. Ambulance	62.50
Req. Fire Service	54.55
Req. Police	41.67
Reporting Injury/Urgency	39.34
Yes	52.94
No	54.17
Others	75.74

Table 2. Classification accuracies of first-pass tags.

5.2 Tag-set improvements

We improved both the DA tagset and the classification framework in our second-pass classification, compared to our initial experiment. We had identified several issues in our first-pass classification:

- We had forced each dialog turn to have a single tag. However, the tags and the dialog turns don't conform to this assumption. For example, the dialog "Yes, my husband has breathing problem. We are at two sixty-one Oak Street"[2] should get 3 tags: "Yes", "Giving-Address", "Requesting-Ambulance".
- Our analysis of the dataset also showed that the initial set of tags are not exhaustive enough to cover the whole range of dialogs required to be translated and conveyed to the dispatcher.

We made several iterations over the tagset to ensure that it is both compact and achieves requisite coverage. The final tag set consists of 67 entries. We manually annotated 59 calls with our new tagset using a web interface. The distribution of the top 20 tags is listed below. The whole list of tags can be found in the NineOneOne project webpage: http://www.cs.cmu.edu/~911/

[2] The dialog is English Translation of *"sí, mi esposo le falta el aire. es acá en el dos sesenta y uno Oak Street"*. It is extracted from the transcription of a CCPD 9-1-1 emergency call, with address modified to protect privacy

Tag (Representation)	Frequency
Yes	227
Giving-Address	133
Giving-Location	113
Giving-Name	107
No	106
Other	94
OK	81
Thank-You	51
Reporting-Conflict	43
Describing-Vehicle	42
Giving-Telephone-Number	40
Hello	36
Reporting-Urgency-Or-Injury	34
Describing-Residence	28
Dont-Know	19
Dont-Understand	16
Giving-Age	15
Goodbye	15
Giving-Medical-Symptoms	14
Requesting-Police	12

Table 3. Distribution of top 20 second-pass tags

The new tagset is hierarchical, which allows us to evaluate the classifier at different levels of the hierarchy, and eventually select the best trade-off between the number of tags and classification accuracy. For example, the first level of tags for reporting incidents includes the five most common incidents, viz, Reporting-Conflict, Reporting-Robbery, Reporting-Traffic-accident, Reporting-Urgency-or-Injury and Reporting-Fire. The second level of tags are used to convey more detailed information about the above incidents (eg. Reporting-Weapons in the case of conflict) or rare incidents (eg. Reporting-Animal-Problem).

5.3 Second-pass classification and Results

We also improved our classification framework to allow multiple tags for a single turn and to easily accomodate any new tags in the future. Our earlier DA classification used a multi-class classifier, as each turn was restricted to have a single tag. To accomodate multiple tags for a single turn, we trained binary classifiers for each tag. All the utterances of the corresponding tag are marked positive examples and the rest are marked as negative examples. Our new data set has 1140 dialog turns and 1331 annotations. Note that the number of annotations is more than the number of labelled turns as each turn may have multiple tags. We report classification accuracies in the following table for each tag based on 10-fold cross-validation:

Tag (Representation)	Accuracy (%)
Yes	87.32
Giving-Address	42.71
Giving-Location	87.32
Giving-Name	42.71
No	37.63
Other	54.98
OK	72.5
Thank-You	41.14
Reporting-Conflict	79.33
Describing-Vehicle	96.82
Giving-Telephone-Number	39.37
Hello	38.79
Reporting-Urgency-Or-Injury	49.8
Describing-Residence	92.75
Dont-Know	41.67
Dont-Understand	36.03
Giving-Age	64.95
Goodbye	87.27
Giving-Medical-Symptoms	47.44
Requesting-Police	79.94

Table 4. Classification accuracies of individual second-pass tags

The average accuracy of the 20 tags is 58.42%. Although multiple classifiers increase the computational complexity during run-time, they are independent of each other, so we can run them in parallel. To ensure the consistency and clarity of the new tag set, we had a second annotator label 39 calls. The inter-coder agreement (Kappa coefficient) between the two annotators is 0.67. This is considered substantial agreement between the annotators, and confirms the consistency of the tag set.

6 Conclusion

The work reported here demonstrates that we can produce Spanish ASR for Spanish emergency calls with reasonable accuracy (76.5%), and classify manual transcriptions of these calls with reasonable accuracy (60.12% on the original tagset,

58.42% on the new, improved tagset). We believe these results are good enough to justify the next phase of research, in which we will develop, user-test, and evaluate a full pilot system. We are also investigating a number of additional techniques to improve the DA classification accuracies. Further we believe that we can design the overall dialog system to ameliorate the inevitable remaining misclassifications, based in part on the confusion matrix of actual errors (Nallasamy et al, 2008). But only actual user tests of a pilot system will allow us to know whether an eventual deployable system is really feasible.

Acknowledgements

This project is funded by NSF Grant No: IIS-0627957 "NineOneOne: Exploratory Research on Recognizing Non-English Speech for Emergency Triage in Disaster Response". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of sponsors.

References

Burges C J C, A tutorial on support vector machines for pattern recognition, In Proc. *Data Mining and Knowledge Discovery*, pp 2(2):955-974, USA, 1998

Finke M, Geutner P, Hild H, Kemp T, Ries K and Westphal M, The Karlsruhe-Verbmobil Speech Recognition Engine, In Proc. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 83-86, Germany, 1997

Frederking R, Rudnicky A, Hogan C and Lenzo K, Interactive Speech Translation in the Diplomat Project, *Machine Translation Journal 15(1-2), Special issue on Spoken Language Translation*, pp. 61-66, USA, 2000

Gao Y, Zhou B, Sarikaya R, Afify M, Kuo H, Zhu W, Deng Y, Prosser C, Zhang W and Besacier L, IBM MASTOR SYSTEM: Multilingual Automatic Speech-to-Speech Translator, In Proc. *First International Workshop on Medical Speech Translation*, pp. 53-56, USA, 2006

Langley C, Domain Action Classification and Argument Parsing for Interlingua-based Spoken Language Translation. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2003

Lavie A, Balducci F, Coletti P, Langley C, Lazzari G, Pianesi F, Taddei L and Waibel A, Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-Commerce Applications,. In Proc. *Human Language Technologies (HLT)*, pp 31-34, USA, 2001

Levin L, Langley C, Lavie A, Gates D, Wallace D and Peterson K, Domain Specific Speech Acts for Spoken Language Translation, In Proc. *4th SIGdial Workshop on Discourse and Dialogue*, pp. 208-217, Japan, 2003

Nallasamy U, Black A, Schultz T and Frederking R, NineOneOne: Recognizing and Classifying Speech for Handling Minority Language Emergency Calls, In Proc. *6th International conference on Language Resources and Evaluation (LREC)*, Morocco, 2008

NineOneOne project webpage [www.cs.cmu.edu/~911]

Schultz T, Westphal M and Waibel A, The GlobalPhone Project: Multilingual LVCSR with JANUS-3, In Proc. *Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop*, pp. 20-27, Czech Republic, 1997

Soltau H, Metze F, F¨ugen C and Waibel A, A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment, In Proc. *IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, Italy, 2001