NineOneOne: Recognizing and Classifying Speech for Handling Minority Language Emergency Calls

Udhyakumar Nallasamy, Alan W Black, Tanja Schultz and Robert Frederking
Language Technologies Institute
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
E-mail: udhay@cmu.edu, {awb,tanja,ref}@cs.cmu.edu

Abstract

In this paper, we describe NineOneOne (9-1-1), 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 towards building the system and the results of our initial experiments.

1. Introduction

In many U.S. localities, when emergency calls are received in languages other than English (primarily Spanish) the dispatching center connects the call to the Language Line human translation service (Language line Services) to translate for them. Though using human translators to assist callers during emergency calls might seem optimal, this scheme actually has serious shortcomings. The process is very slow, especially in starting up, and the translators are unfamiliar with the task, resulting in very poor quality service.

We are applying and adapting speech translation technology to the domain of 9-1-1 emergency dispatching. (The standard emergency telephone number in the United States is "9-1-1".) The 9-1-1 domain has many research challenges, but we believe it is also a feasible domain for a real-world speech translation application. The domain is challenging because it requires real-time operation and the recognition and translation of stressed telephone-quality speech in multiple dialects; it is still feasible because we have significant in-domain data, there are strong vocabulary and task constraints, and perfection is not required. This domain also has clear, significant social value, addressing the chronic shortage of translation of Spanish (and in larger cities, a large number of other languages) at U.S. emergency dispatch centers. While we currently are working on Spanish/English, the approaches we use are largely language-independent.

We expect that this project will eventually lead to an actually deployed system that performs full automatic speech recognition (ASR) and targeted, classification-based machine translation (MT). This would be an important scientific result, validating for the first time the hypothesis that speech can be adequately recognized and translated for real-world use. Our initial work is aimed at demonstrating that we can produce ASR and utterance classification of sufficient quality to allow the development of such a practical, limited-domain system. The results reported here achieve the goal of justifying further work in building, user-testing, and evaluating a full pilot system.

2. Domain Overview

When someone dials 9-1-1 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. The major initial decision is whether to send police, fire or medical personnel; the appropriate units are dispatched as soon as this decision is made. While the responding unit is en route, the dispatcher attempts to elicit more details about the emergency from the caller. The additional details are intended to help the responding unit prepare for the situation that they will encounter. For example, the police want to know the level of violence involved in advance. Will they be walking into an armed confrontation, or is someone reporting a crime that occurred yesterday? Another important issue is the exact location of the emergency; although the 9-1-1 equipment 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 (and Voice-over-IP) do not generally provide 9-1-1 location information. As the dispatcher elicits more information about the emergency, they radio it to the responding unit.

3. System Architecture

Full reliable speech-to-speech translation is still beyond our capabilities, especially for real-time human-directed conversation. However in many applications, full translation is not actually required, and a more limited form is adequate (Stallard et al, 2007; Gao et al, 2006). Based on the domain's characteristics, we are following a highly asymmetrical design for the eventual full system (Frederking et al, 2000), see Figure 1. The dispatcher is already seated at a workstation, and we intend to keep them "in the loop", for both technical and social reasons. In the dispatcher-to-caller direction, we can work with text and menus, so 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.

In the final system, we will adapt an approach that was demonstrated in the joint NSF/EU "Nespole!" project (Lavie et al, 2001): Domain Action (DA) classification (Levin et al, 2003; Langley 2003). We use the term DA to refer to the combination of a general Speech Act with domain-specific concepts. DAs capture speaker intention in a limited-domain system, rather than detailed literal meaning, and thus help the MT system cope with ungrammaticality and ASR errors, by ignoring out-of-domain speech and focusing on recognizing the main intent and crucial details of each utterance. Example DAs in this domain might be Request-Ambulance ("I need an ambulance!") or Giving-Address ("I live at 2635 Rodeo Drive"). Once we have classified the utterance into a DA, the next step in the eventual full system will be to identify and translate just the arguments of the DA. An example output argument structure for the second DA above might be (address-number="2635", address-street="Rodeo Drive").

In this initial work, we only seek to demonstrate that we can carry out utterance classification on ASR output of sufficient quality to support this approach to MT. The actual MT system will be developed in a follow-on project.

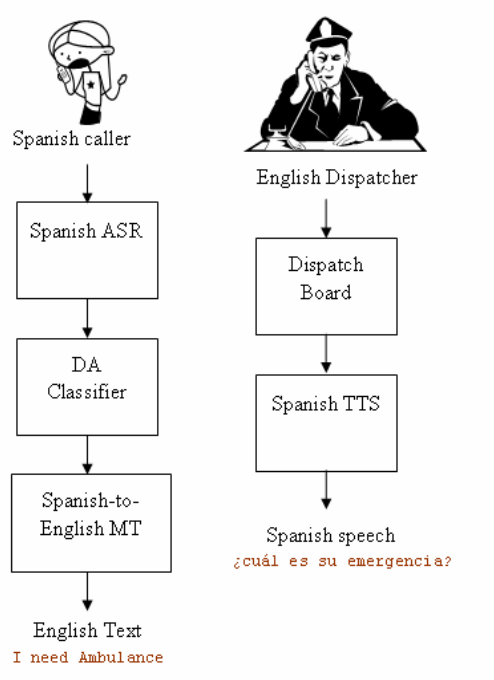

Figure 1: System Flow: a) Caller to Dispatcher and 2) Dispatcher to Caller

4. Automatic Speech Recognizer (ASR)

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 between 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, sub-phonetically tied semi-continuous 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 32 dimensional Mel-scale cepstral coefficients and applies Linear Discriminant Analysis (LDA) calculated from the training data. 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 don't 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. 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.

5. Utterance Classification

As described in Section 3 above, the speech recognizer output needs to be classified into domain-specific actions (DAs) for eventual translation, as well as dialogue control. So as an initial evaluation, experiments were carried out on classifying manual audio transcriptions of the 9-1-1 calls used to train the speech recognizer. We extracted all the dialog turns in the transcriptions between the human translator and the caller and labelled them according to the DA of the caller. The following tags were used to label each turn of interaction between the interpreter and the caller: Giving-Name, Giving-Address, Giving-Telephone Number, Requesting-Medical-Assistance, Requesting -Fire-Service, Requesting-Police, Reporting-urgency-or -injury, Yes, and No. The tags "Yes" and "No" are labels for the answers to yes-no questions. An additional tag called "Others" is used to label all other cases. The

database has a total of 845 labelled turns. The distribution of different tags in the corpus is shown in Table 1 below.

Tag (Representation)	Frequency
Giving Name (GN)	80
Giving Address (GA)	118
Giving Phone number (GP)	29
Requesting Ambulance (RA)	8
Requesting Fire Service (RF)	11
Requesting Police (RP)	24
Reporting Injury/Urgency (RI)	61
Yes (Y)	119
No (N)	24
Others (O)	371

Table 1: Distribution of tags in the corpus.

We used the Support Vector Machines (SVM) (Burges, 1998) implementation in the WEKA Machine learning toolkit (Garner, 1995) to classify the turns based on Domain Acts. Simple bag-of-words binary features are used for classification. Individual accuracies for each tag are given in Table 2 below.

Tag	Accuracy (%)	Accuracy (without "Others" tag in %)
Giving Name	57.50	67.6
Giving Address	38.98	63.0
Giving Phone number	48.28	63.6
Req. Ambulance	62.50	83.3
Req. Fire Service	54.55	75.0
Req. Police	41.67	62.5
Reporting Injury/Urgency	39.34	72.7
Yes	52.94	71.6
No	54.17	81.2
Others	75.74	----

Table 2: Classification accuracy of individual tags.

The overall accuracy of the classifier on 10-fold cross-validation is 60.12%. However, if we leave out all misclassifications of tags as "Others", the overall accuracy of the classifier improves to 68.8%, as shown in the last column of Table 2. This analysis is interesting due to eventual dialogue system design issues. If the full system classifies an utterance with the "Others" tag, it will need to prompt the caller with more specific questions (e.g. "Do you need an ambulance?") to understand their intent. Thus the accuracy without "Others" more accurately reflects the expected performance of the eventual full system.

The classification in this initial evaluation was done on manual transcriptions. When we apply the same techniques to automatic transcriptions (ASR output), the classifier accuracy falls to 40% (or 49% when excluding the "Others" category). This performance will probably need to be improved in order to produce a working pilot system. We intend to explore a number of potential classifier improvements, including:

- Using LM perplexity as a feature
- Using word classes as a feature
- Manually building special domain-specific feature recognizers, such as a "telephone number recognizer"
- Using synonyms from Spanish EuroWordNet to smooth the match between utterances
- Using discourse context to bias classification

These techniques would be in addition to the more mundane improvements we hope to have from acquiring, transcribing and labeling more data (currently in progress), improving the utterance tag-set (also in progress), and further improving the ASR.

Another possibility is to use a second stage of classification on highly-confusable classes (Liu et al, 2003). Note in the confusion matrix (Table 3) that the matrix is fairly sparse; the misclassifications only occur between certain classes. We could train a set of classifiers on only these classes, to separate them.

Tag	GN	GA	GP	RA	RF	RP	IU	Y	N	O
GN	46	9	-	-	-	1	3	9	-	12
GA	9	46	-	-	2	-	3	13	-	45
GP	-	8	14	-	-	-	-	-	-	7
RA	-	1	-	5	-	-	-	-	-	2
RF	-	-	-	-	6	-	2	-	-	3
RP	1	-	-	-	-	10	5	-	-	8
IU	3	-	-	6	-	-	24	-	-	28
Y	9	11	5	-	-	-	-	63	-	31
N	-	-	3	-	-	-	-	-	13	8
O	12	27	-	2	3	4	7	25	10	281

Table 3: Confusion matrix between classes

It must also be noted that the current evaluation utilized manual segmentation of recordings by speaker. The final system will be integrated with the 9-1-1 answering and dispatching unit which has access to the uninterrupted caller channel of the full-duplex telephone line, thus avoiding segmentation. However, we will need Voice Activity Detection (VAD) to identify endpoints for ASR.

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 (68.8% on human transcripts, ignoring the "Others" category). We believe these results are clearly good enough to justify the next phase of research, in which we will develop, user-test, and evaluate a full pilot system.

Only actual user tests of a pilot system will allow us to know whether an eventual deployable system is actually feasible.

7. 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.

8. 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

Garner S R, WEKA: The Waikato environment for knowledge analysis, In Proc. *New Zealand Computer Science Research Students Conference*, pp. 57-64, New Zealand, 1995

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

Language Line Services [http://www.languageline.com]

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

Liu Y, Carbonell J and Jin R, A pairwise ensemble approach for accurate genre classification. In Proc. *14th European Conference on Machine Learning (ECML)*, Croatia, 2003

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

Stallard D, Choi F, Kao C, Krstovski K, Natarajan P, Prasad R, Saleem S and Subramanian S, The BBN 2007 Displayless English/Iraqi Speech-to-Speech Translation System, In Proc. *International Conference on Spoken Language Processing (Interspeech)*, Belgium, 2007