A COMPARISON OF LANGUAGE PROCESSING TECHNIQUES FOR A CONSTRAINED SPEECH TRANSLATION SYSTEM

Mike Lincoln and Stephen Cox

School of Information Systems, University of East Anglia, Norwich, NR47TJ

ABSTRACT

A system designed to allow Post Office counter clerks to communicate with deaf customers by translating speech into sign language is described. The system uses approximately 370 pre-stored phrases which may be signed to the customer using a specially designed avatar. The clerk is unable to memorise this number of phrases and therefore the system attempts to map from their input speech to the semantically equivalent pre-stored phrase. We describe a number of language processing techniques developed to perform the mapping, and give results obtained using alternative formulations of the phrases from a number of speakers. We then give results for recognised speech input and show how misrecognitions effect the mapping system. Best performance is obtained using a mapping system based on an entropy weighted, vector based distance measure between the test phrase and each of the signed phrases.

1. INTRODUCTION

TESSA (The Text and Sign Support Assistant) is a system designed to translate the spoken language of a Post Office counter clerk into British sign language (BSL) to assist in the completion of a transaction between the clerk and a deaf customer [1]. A priori, it may seem that simply displaying the text spoken by the counter clerk would be sufficient for a system designed to aid deaf people. However, for many who have been profoundly deaf from an early age, BSL is their first language. Such people learn to read and write English as a second language and often have below average reading skills, preferring instead to use BSL as their primary communication medium. BSL is a fully developed language, with a syntactic structure and grammar which is quite different from that of spoken languages, and less well understood. No standard phonetic representation of the components of the language exists and so creating a general purpose translation system from spoken to signed language is a formidable research problem. While some progress has been made on such translation systems in recent years [2], they are yet to be robust or general enough for use in a system such as TESSA. Instead we take a different approach to the language translation problem, based on the fact that the domain in which TESSA operates (that of Post Office transactions) is highly constrained in both its scope and topic. Recent research on "formulaic language" [3] has shown that under such conditions, cross langauge communication is possible using a limited set of predefined phrases relevant to the domain. We therefore avoid the problem of having to develop a general purpose translation system by selecting a number of phrases which may be signed by the system, and mapping from the clerk's input speech to the most relevant of the phrases which are available to the system. This paper is concerned with the system used to map from the clerk's speech to the signed phrases and describes a number of different techniques which have been investigated to perform the task.

Canonical Phrase	Alternative
P1: Where is it going to	Where are you sending this
	Where will it go to
	Where is the letter going to
	Where is the destination
	Where is the item going
P2: Where did you obtain this	Where did this come from
	Where did you get this from
	Where did you get this from
	Where did you get this from
	Where did you get it from

Table 1. Phrases and alternatives

2. PROBLEM DESCRIPTION AND DATA

The output from TESSA consists of different signed phrases, motion captured from a human signer and replayed on a specially developed avatar. The phrases were chosen from the analysis of recordings of transactions performed in 3 Post Offices around the UK, in total about 16 hours of business. We estimate that the 370 phrases identified from the analysis cover approximately 90% of transactions. These 370 phrases are referred to as the 'canonical phrases' and it was these which were used as prompts for the signers during the motion capture of the sign sequences. In order to obtain data to test the phrase mapping systems, we took a subset of 155 of the canonical phrases and asked 5 volunteers to each give an alternative but semantically equivalent way of expressing each phrase. These data were used to test the robustness of the mapping algorithms to variations in the vocabulary and syntax used by the speakers to express the canonical phrases. Table 1 shows two examples of the phrases and their alternatives.

Each volunteer was also recorded speaking each of their alternative phrases and these were subsequently transcribed by a speech recogniser and used to investigate the effects of recognition errors on the effectiveness of the mapping techniques. The recogniser was trained using HTK [4] on the speaker-independent training set of the WSJCAM0 British English speech database (92 talkers, $\sim$ 90 utterances per speaker) and consisted of 3500 HMM states with 8 Gaussian mixture components per state. A bigram SLM trained on all the training set phrases was used. Recognition accuracy results for the system were 75.1% correct and 49.8% accurate. Speaker adaptation techniques are well known to produce a significant increase in recognition accuracy, particularly when, as in this case, the training and testing data are recorded in different acoustic

C1: Where is it going to
C2: How much is the item worth
C3: Does it contain anything valuable
C4: What country are you travelling to
C5: Do you want first or second class stamps
recognised speech: What country is it going to

Word	C1	c2	c3	c4	c5
what	0	0	0	1	0
country	0	0	0	1	0
is	1	1	0	0	0
it	1	0	1	0	0
going	1	0	0	0	0
to	1	0	0	1	0
$\sum$	4	1	1	3	0

Output Phrase = C1: Where is it going to

Table 2. Phrase mapping using Co-occurrence of words in input speech and canonical phrases (Method 1). C1 - C5 are the Canonical Phrases

environments. Unfortunately no adaptation data was available and as such the recognition performance is less than optimal. A full description of the recognition system can be found in [5].

Because there was insufficient data from each volunteer to divide into separate training and testing sets, we adopted a leave-one-out approach to training. For testing on speaker S_k, the algorithms were trained on the canonical phrases and their alternatives from all speakers except speaker S_k. Testing data was either

1. The text transcriptions of the alternative phrase from speaker S_k
2. The recognised output of speaker S_k's recorded alternative phrases.

3. PHRASE MAPPING ALGORITHMS

3.1. Method1 - Co-Occurring words without alternatives

The first method for performing the phrase mapping is a simple system which uses only the canonical phrases as training data, and not the alternative phrases provided by the volunteers. Table 2 shows an example of the phrase mapping system with 5 canonical phrases. The number of words occurring in both the input speech and each of the Canonical phrases is calculated. The phrase or phrases (in the event of a tie) with the highest number of co-occurrences is identified as the candidate phrase. This simple approach gives us some insight into how constrained the problem is — if such a system were able to achieve a high mapping accuracy, then the phrases are sufficiently different from each other that a more complex algorithm is unlikely to provide significant increases in performance.

3.2. Method2 - Co-occurring words with alternatives

The training data for Method 1 does not include the alternative ways of saying the phrase and hence, if the users use a large and diverse vocabulary for expressing the phrase, the system will often fail. The second method is identical to the first except that co-occurring words in the test phrase and both the canonical and alternative phrases are calculated. It should be noted that although alternatives were obtained from 5 subjects, in many cases the alternatives were identical from the majority or sometimes from all the candidates, effectively reducing the size of the training set. Again the phrase or phrases with the highest number of co-occurring words is returned as being correct. It was expected that this system would alleviate the problem of diverse vocabularies resulting in incorrect mapping for certain phrases.

3.3. Method 3 - Entropy weighted Co-occurring words

While the previous method takes into account the alternative pronunciations, it has the disadvantage of giving equal weight to every word. It would be expected that some words are of more importance than others in performing the phrase mapping and should be given a greater weighting accordingly, whilst words which occur in a large number of phrases and therefore have little discriminatory power (e.g. function words) should be given a lower weighting. We require for each word to be assigned a 'score' which takes into account the frequency with which the word occurs in each phrase. Each word was therefore given a score based on the conditional entropy of the word given the phrase. The score is calculated as follows: If $W(i,j)$ is a count of the number of times word i occurs in phrase j and its alternatives, then, if there are N canonical phrases, the probability of observing phrase p_j given that word w_i is observed is

$$Pr(P_j|w_i) = \frac{W(i,j)}{\sum_{k=1}^{N} W(i,k)} \quad (1)$$

and the conditional entropy of the phrases given word w_i is

$$H(w_i) = -\sum_{k=1}^{N} Pr(p_k|w_i) log_2 Pr(p_k|w_i) \quad (2)$$

$H(w_i)$ has a maximum value of $log_2 N$ when the $Pr(p_j|w_i)$ are equiprobable and a minimum value of 0 when w_i occurs in only one phrase. We define a score, E_i, for each word as

$$E_i = 1 - \frac{H(w_i)}{log_2 N} \quad (3)$$

E_i is in the range $0 \leq E_i \leq 1$. E_i has a value of zero when w_i occurs with equal probability in every phrase, and a value of one when w_i occurs in only one phrase. When the clerk's speech is recognised, phrases are ranked according to the sum of the scores of words occurring in both the recognised phrase and the canonical phrase and as before, the highest ranking phrase or phrases are returned as the most likely.

3.4. Method 4 - Vector based classification with entropy weighting

Recently, there has been considerable interest in automatic call routing, where a spoken request from a user is automatically directed to the correct department or extension [6, 7]. If we regard each canonical phrase as a 'destination' and the input phrase as a "query", then the task of identifying the correct phrase is equivalent to the telephone call routing task. The fourth phrase mapping system still uses an entropy-based measure to weight each word, but employs a vector-based approach to the classification. Such an approach which has been shown to give superior performance to other systems in the call-routing task [6]. For mapping, each

Word	P1	P2
where	6	6
is	4	0
it	2	1
going	3	0
to	2	0
are	1	0
you	1	4
send	1	0
this	1	4
will	1	0
go	1	0
the	3	0
letter	1	0
destination	1	0
item	1	0
did	0	6
obtain	0	1
come	0	1
from	0	6
get	0	4

Table 3. A co-occurrence matrix made from two canonical phrases and their alternatives

phrase is represented as a vector and mapping to the "semantically equivalent" canonical phrase is done by estimating the distance between the input speech vector and those representing the canonical phrases.

Classification is performed as follows. A co-occurrence matrix is generated where each column represents a phrase and each row represent a word in a phrase. To illustrate this, the co-occurrence table for the example phrases in Table 1 is shown in Table 3.

As has been noted, simply using the counts of the words is not good for classification since the words occur with very different frequencies. Again, we choose to use a measure based on the conditional entropy of the phrase given the word. However, instead of simply weighting each element of $W(i,j)$ by E_i as given by equation 4 we use

$$W(i,j) \rightarrow \log\left(1 + \frac{W(i,j)}{O_j}\right) * E_i \qquad (4)$$

In Equation 4, O_j is the number of different words in phrase P_j. If O_j is small, the term $\log\left(1 + \frac{W(i,j)}{O_j}\right)$ is large, so a word associated with a phrase that has few other words associated with it, has a large weight. Conversely, if the same word were associated with a document that had many other words associated with it, the weight of this term would be smaller. This weighting was proposed by Bellegarda [8].

To perform the classification we augment the co-occurrence matrix of Table 3 with an additional column of entries for the words in the recognised phrase. We then calculate the terms E_i and O_j and weight the augmented matrix. Each column in the matrix is now regarded as a vector and the distance between the vector representing the input phrase and each of the canonical phrases is calculated, to determine which canonical phrase is 'closest' to the input speech. The dot product between the two vectors was found to give better performance than the Euclidean, or normalised Euclidean distance and is used to calculate the distance measure.

	% correctly mapped					
Method	1	2	3	4	5-2Gram	5-3Gram
Text	82.2	93.5	95.2	97.5	98.6	98.5
Recognised	65.9	72.4	78.4	82.2	81.6	81.7

Table 4. Number of correctly mapped phrases when number of false positives = 7%

3.5. Method 5 - Vector based classification with word n-gram's

The final method investigated is an extension to Method 4 which incorporates word order within the phrase a feature which has, until now, been ignored. During its construction, the co-occurrence table is augmented with rows representing word n-grams from the training phrases. For example, the phrase 'where is it going to' would add entries 'where is', 'is it' and 'it going' to the co-occurrence matrix as well as the individual words in the phrase. These n-grams are then used during the classification in exactly the same way as single words. However their occurrence will be less frequent and, as such, they will have a higher entropy score.

4. EXPERIMENTAL DETAILS AND RESULTS

The methods were tested on the text of the alternative phrases from each of the five volunteers. Each method returned the top 1 to 5 scoring phrases and from this the percentage of correct positives (that is phrases returned which were the correct phrase as a percentage of the total number of possible correct phrases) and false positives (ie the total number of phrases returned minus the number correct, as a percentage of the total number of phrases which could possibly have been returned) were calculated. It should be noted that, particularly in the case of methods 1 and 2 for which each phrase is given an integer score, many phrases have the same score and as such there may be many equally likely phrases in the Top N. Figure 1 shows ROC curves for methods 1 to 4. Method 1 is significantly worse than any of the other methods, as would be expected since it has only one sixth of the training data available. Method 2 improves on this. However the integer scoring method results in large numbers of false positives, ultimately limiting the usefulness of the technique. Method 3 provides a further increase in performance and also avoids the large number of false positives by using a non-integer score to rank the phrases. The improved classification technique of Method 4 increases the accuracy further still and, as shown in Figure 4, augmenting the system with N-grams increases the accuracy by approximately 1% with no increase in false positives. Table 4 gives the accuracy of each of the techniques when the number of false positives is 7%. This is approximately equivalent to the system returning the top 10 candidate phrases - the clerk would then be expected to select the correct phrase from this list.

Figure 3 shows the effect of recognition errors on the performance of each of the techniques. The same pattern in the ranking of the methods exists and the number of false positives is largely unchanged. However, the number of correctly mapped phrases has dropped by approximately 15-20% in all cases simply because of the number of erroneous words in the input. Figure 4 shows that using bigrams and trigrams to augment the co-occurrence vectors in Method 5 has virtually no effect on the accuracy. This may be because the recognition errors tend to be evenly spread through the text rather than occurring in clusters, and as such many pairs of words contain recognition errors.

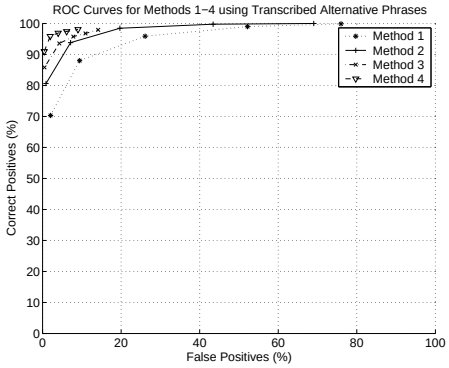

Fig. 1. Mapping performance for Methods 1–4 - Text Transcriptions.

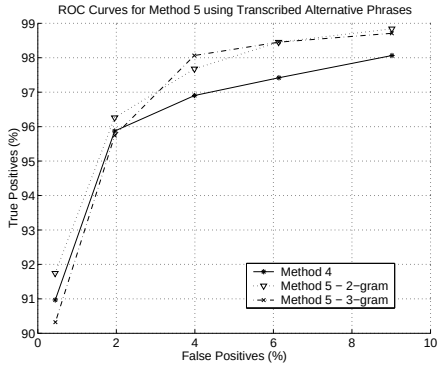

Fig. 2. Mapping performance for Method 5 - Text Transcriptions

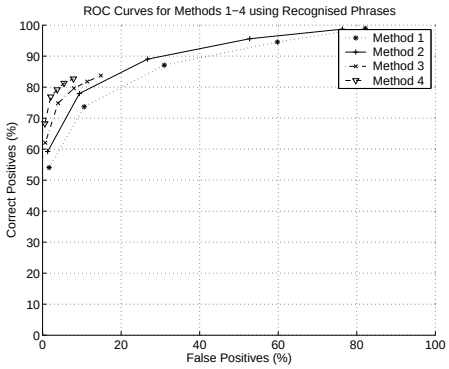

Fig. 3. Mapping performance for Methods 1–4 - Recognised Speech.

5. DISCUSSION AND FUTURE WORK

We have presented a number of different techniques for mapping from speech to semantically equivalent pre-recorded signed phrases in the domain of Post Office transactions. We have shown that the best performance is obtained using a weighting on the words in the phrase based on their entropy, and a mapping using a vector based distance measure between the input phrase and the signed phrases. However extending this technique to include word n-grams in the scoring procedure provided little increase in accuracy. Best performance on text transcriptions of users alternative formulations of the phrases was 98.6% when the number of false positives was 7%.

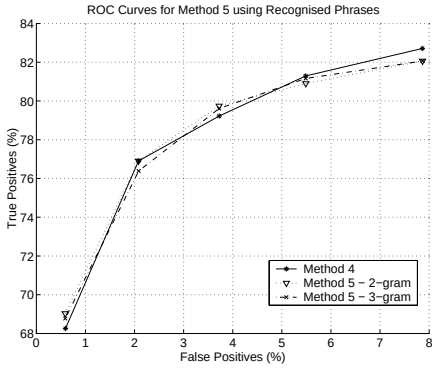

Fig. 4. Mapping performance for Method 5 - Recognised Speech

Mapping performance on recognised speech was significantly worse than using text transcriptions because of the large number of recognition errors in the input. Best phrase retrieval performance was 82.2% for 7% false positives. It should be noted however that the recognition performance was far from optimal due largely to the lack of speaker adaptation. Currently the system has no means of dealing with 'out of vocabulary' words and future work could include investigating methods to include OOV words in the classification process.

6. REFERENCES

[1] S.J. Cox et al, "TESSA, a system to aid communication with deaf people," in *Proc. ASSETS 2002, Fifth International ACM SIGCAPH Conf. on Assistive Technologies.)*, July 2002.

[2] E. Safar and I. Marshall, "The architecture of an english-text-to-sign-languages translation system," in *Recent Advances in Natural Language Processing (RANLP)*, G. Angelova, Ed., pp. pp223–228. Tzigov Chark, Bulgaria, 2001.

[3] A. Wray, "The functions of formulaic language: an integrated model.," *Language and Communication*, vol. 20, no. 1, pp. 1–28, 2000.

[4] J. Jansen, J. Odell, D. Ollason, and P. Woodland, *The HTK book*, Entropic Research Laboratories Inc., 1996.

[5] S.J. Cox, "Speech and language processing for a constrained speech translation system," in *Proc. Int. Conf. on Spoken Language Processing*, September 2002.

[6] J Chu-Carroll and R Carpenter, "Vector-based natural language call-routing," *Computational Linguistics*, vol. 25, no. 3, pp. 361–388, 1999.

[7] S.J. Cox and B. Shahshahani, "A comparison of some different techniques for vector based call-routing," in *Proc. 7th European Conf. on Speech Communication and Technology*, September 2001.

[8] J.R. Bellegarda, "A multispan language modeling framework for large vocabulary speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 456–467, September 1998.