Improving speech recognition by utilizing domain knowledge and confidence measures

Pascal Wiggers, Leon J. M. Rothkrantz

Data and Knowledge Systems Group
Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands
{p.wiggers, l.j.m.rothkrantz}@its.tudelft.nl

Abstract. In speech recognition domain knowledge is usually implemented by training specialized acoustic and language models. This requires large amounts of training data for the domain. When such data is not accessible there often still is external knowledge, obtained through other means, available that might be used to constrain the search for likely utterances. This paper presents a number of methods to exploit such knowledge; an adaptive language model and a lattice rescoring approach based on Bayesian updating. To decide whether external knowledge is applicable a word level confidence measures is implemented. As a special case of the general problem station-to-station travel frequencies are considered to improve recognition accuracy in a train table dialog system. Experiments are conduced to test and compare the different techniques.

1 Introduction

Typical speech recognizers have to face the daunting task of unraveling the complex speech signal armed with nothing but a set of acoustic models and a very limited understanding of grammar embodied in a statistical language model, usually a bigram or trigram model. However, in many cases there is much more knowledge available to constrain the search for correct utterances. On one hand there is a large body of linguistic knowledge, which is for example successfully exploited in the structured language model of [1] while on the other hand the context of use or information obtained form other program components may also provide cues about the words to come. This is especially the case in systems that operate on a limited domain, like dialogue management systems. Over the past few years such a dialogue system for train table information of the Dutch railways has been developed within our group [4]. In this paper we explore some possibilities to use domain knowledge within this system. The domain knowledge we concentrate on are the identities of the stations involved. Analysis of train table dialogues revealed that there is a strong correlation between departure and arrival stations. Knowing one of the two stations provides strong indications concerning the identity of the other one.

Given such knowledge, or for that matter domain knowledge in general, there are two questions that have to be answered. First, it has to be determined at what stage in the recognition process this data should best be employed. To answer this question we

experimented with two different methods. The first approach, which has been taken by a number of authors to take advantage of domain knowledge, is by updating the language model probabilities [8]. In this case the search process is directly influenced. The other approach uses the connection information in a post-processing step.

The second question that has to be answered is in what cases the external knowledge should be used in the first place. Our baseline system recognizes over 85% of all station names correctly. So, in many cases the use of additional knowledge is not necessary and might actually cause more harm than good. This is especially true as a number of stations (typically the major cities) have high connection frequencies with almost all other stations. Thus, the utilization of frequency information should be limited to those cases where there is uncertainty in the recognizer output. For this purpose a confidence measure based on word posterior probabilities is implemented.

2 Data analysis

During the development of the dialogue system a large data set of telephone conversations between users and operators has been collected and transcribed.
The set, called the OVR log files, covers the conversations of an entire year (October 1995 – October 1996). The total number of references to stations in questions regarding station-to-station queries is almost 13 millions. Analysis of this set revealed the following facts [5]:

- The connections from and to the first 50 stations (out of 377 stations) make up 50% of all connections asked for.
- Connections are highly dependent on the stations involved.
- There is a strong inverse correlation between the distance to travel and the frequency of these journeys

Given the dependencies found in the log files it may be clear that this information can be useful in recognizing the correct station name. Knowledge of either the departure or arrival station can help to restrict the search for the other station name considerably.

Figure 1 illustrates the general idea. It shows n-best lists for the departure and arrival stations. Normally the first best option from both lists would be taken. However the connection frequencies, depicted by the thickness of the lines in figure 1, suggest that the second best option for the departure station is very likely given the first best arrival station.

As there are a number of stations that always have high frequencies, while on the other hand the majority of the connections have relatively low frequencies, relying too much on the connection frequencies does not seem very wise. This may lead to overall improvement at the expenses of small stations. Therefore, a combination of the acoustic evidence for the alternatives and the connection frequencies should be used to decide which station pair to choose, this is where the confidence measures come into play.

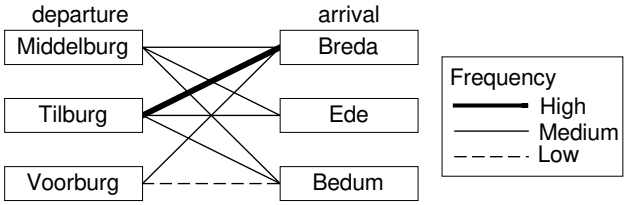

Fig. 1. Using connection frequencies to select best combination of station names

3 Confidence measures

The confidence measures are based on a lattice output by the recognizer. The main feature used is the word posterior probability, i.e. the probability of a word in a certain position given the observation sequence. As speech recognizers try to maximize the posterior probability over an entire utterance the word level posteriors are usually not known, but they can efficiently be approximated using a variation of the forward-backward algorithm on the output lattice [3]. First the posterior probabilities of all word links in the lattice have to be calculated. These can be defined as the sum of the probabilities of all paths passing through the link normalized by the probability of the complete observation sequence. The latter is approximated by the sum of all paths through the network. The word level posterior is now the sum of all link posteriors corresponding to this word. However, deciding which links correspond to which word is not trivial. We chose to proceed along the lines of [2] by clustering the words in the lattice based on overlap in time. Links in the same cluster that correspond to the same word are combined. An example of the resulting network is shown in figure 2.

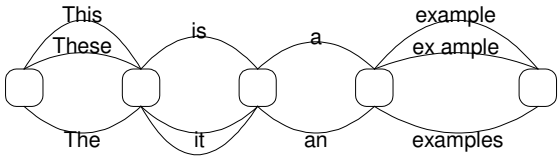

Fig. 2. A confusion network

This type of network is often referred to as confusion network and explicitly shows which words in the lattice can be considered competing alternatives to each other. Our procedure slightly differs from the ones described in [2,7] as we allow successive words to be in the same cluster, this may happen when two smaller words correspond to one longer word (e.g. ‘ex’ and ‘ample’ to ‘example’). For further processing these sequences in a cluster are seen as a single entity.

Aside from the posterior probability a number of other confidence features is determined for each cluster:

- *Posterior drop*: The quotient of the posterior probabilities of the first and second best option. If this value is close to one the second best alternative is almost as good as the first best, which may indicate low confidence.
- *N-best position*: The first hypothesis the word occurs in. This may differ from the word having the highest posterior probability.
- *Relative N-best position*: The number of consecutive hypothesis the word occurs in. This is a measure of the stability of a word.
- Number of words in a cluster: Many alternatives may imply lower confidence.
- *Number of words on a cluster arc*: As the alternatives in a cluster can actually be word sequences, this number may be larger than one. This usually means that a longer word has been recognized as a sequence of smaller words.
- *Language model score*: Words having a higher language model score may be better trained and thus better recognized.
- *Confidence value of the previous cluster*: Recognition errors tend to occur in bursts.

These features were selected as the most discriminative subset from a larger set of features, including the number of frames, the normalized acoustic score, the number of links corresponding to a word and the average acoustic score drop. The latter is defined as the average of the frame-wise differences of the acoustic scores of the first and second best results, this is similar to the measure defined in [9].

Inspection of the intercorrelation of the features and the correlation of each feature with the output as well as the information gain of each feature showed that the combination of posterior probability, n-best position and relative n-best position does most of the job.

To see which features should be added to get the most optimal results a greedy hillclimbing search through feature space was performed, for each subset a decision tree classifier was built and tested [6]. For combination of the features in the most promising subsets we experimented with several classifier types including neural networks and linear regression trees but we found that there are no significant performance differences. For the final system a linear regression tree was chosen. Table 1 shows the classification results of the confidence measure on the phonetically rich sentences of the evaluation part of the Dutch Polyphone database, the word error rate of the recognizer on this set is 9%.

Table 1. Detailed results confidence measure

Correctly classified instances			94.22%	
Class	TP rate	FP rate	Precision	Recall
CORRECT	98.12%	46.50%	98.12%	95.53%
INCORRECT	53.50%	01.88%	53.50%	73.79%

4 Marginal frequencies

The marginal departure and arrival frequencies, that is, the probability that a user wants to travel to or from a certain station can directly be used in the language model.

To do so, we trained a standard bigram language model but replaced all station names with the respective labels *from_station* and *to_station*, assuming that the words preceding a name do not depend on that particular name. Two classes were used instead of a general station class as data analysis also revealed that frequencies are not completely symmetrical. For example it was found that there are much more requests for journeys to the train station of Amsterdam Airport than for traveling from the airport to another station. A likely explanation may be that people often do not know when exactly they will be back at the station. The probabilities within these classes correspond to the marginal frequencies; these are combined with the bigram class probabilities according to:

$$P(w_i|w_{i-1}) = P(g(w_i)|g(w_{i-1}))P(w_i|g(w_i)) \quad (1)$$

Taking advantage of the connection frequencies is less straightforward as these do not fit very well in an n-gram approach, since the station names may be separated by a large, unknown number of words, be in different utterances or they may be obtained from other sources, for example a pointing device in a multi-modal system. As we are ultimately interested in the more general concept of using external knowledge in speech recognition these possibilities should be taken into account.

We considered two basic approaches: an early integration approach, where the connection frequencies are used directly in the language model and a late integration approach where the connection frequencies are used in a post-processing step on the results produced by the recognizer.

5 Early use of connection frequencies

In this case the first name is recognized using the class based language model described above. If the confidence value of this name is above a predefined threshold τ the name is assumed to be correct and the language model probabilities are updated using the connection frequencies from this station to all other stations. This is implemented by a weighted interpolation of the conditional probabilities derived from the frequency table and the language model probabilities:

$$P(w_i|w_{i-1}) = \lambda P_{lm}(w_i|w_{i-1}) + (1-\lambda)P_{conn}(w_i) \quad (2)$$

In case of low confidence for the first name no assumptions can be made about likely connections and the standard language model is also used for recognition of the second name.

The drawback of this approach is that uncertainty about the first name is not resolved, even though the second name may be helpful here. In addition, the method cannot handle the case where both names are mentioned in the same utterance. As a solution a second recognition pass was introduced for the first name using a language model updated based on knowledge of the second name. This is a computationally rather expensive method, a more efficient, albeit probably less accurate method, would be to rescore the original recognition results with the updated language model. As we are only interested in station names and not in whatever other words are

recognized it might be attractive to take this idea one step further and use the connection frequencies in a post-processing step, which has the additional advantage that the influence of the different knowledge sources can be better controlled.

6 Using frequencies for rescoring

The rescoring approach uses the normal language model containing the marginal frequencies to produce n-best lattices for both station names, which are transformed to confusion networks. The clusters containing station names are located and the connection frequencies are used to pick the best pair of names. This method is particularly prone to introduction of errors as frequent connections will always be chosen, thus as before the confidence measures are used to limit the use of frequency information to those cases where there remains uncertainty about one of the names. In those cases where frequency information is utilized it is combined with the acoustic evidence using a Bayesian updating approach. The frequency table is used to obtain the joint probability of a connection from departure station S1 to arrival station S2. The word level posterior probabilities, which are already calculated during confidence measure recognition, are used to represent the probability that a station name was uttered given the corresponding observation (sub)sequence. Assuming statistical independence of the second station name and the first observation sequence given the first name this can be written as:

$$P(S_1|S_2O_1) = \frac{P(S_1S_2)P(S_1|O_1)}{P(S_1)\sum_{S_1} P(S_2|S_1)P(S_1|O_2)} \quad (3)$$

As we are only interested in relative scores the summation can be omitted. The formula has an intuitive interpretation as the overall score will be higher when a name is more likely to be pronounced and is likely to occur together with the other station name, while stations with a small marginal probability, that is the rarer stations are preferred to stations that have a high frequency in general.

Although we now have a nice theoretic framework with a clear interpretation, one may wonder if it really is an appropriate choice in this case. First of all the different pieces of evidence do not seem to be of equal importance, especially the role of the marginal probabilities is rather small in practice. Furthermore, as word lattices cover only part of the entire search space the posterior probabilities are just an approximation of the real posteriors; in general they tend to overestimate the probabilities of the best alternatives. To smooth the distribution of the posterior estimates of competing alternatives we may also consider other features than the posterior probability and combine them into a single value, effectively implementing a confidence measure for each alternative. This was done by training a linear regression classifier that uses features that relate directly to the alternative under consideration: the posterior probability, the n-best position, the relative n-best position, the language model score and the number of frames.

As each classifier works independently form the other alternatives the resulting values can no longer be seen as probabilities, so the Bayesian framework is no longer

applicable. However, the reasoning behind the formula given above still stands. Therefore we resided to a similar although somewhat less formal approach. The acoustic evidence for the alternatives of the stations, represented by their confidence measures and the connections frequencies are combined using a weighted geometric mean, which has the desired effect that the overall score is high when its individual items are high, while it is rather insensitive to high values of a single attribute.

$$Score(S_1S_2) = C(S_1|O_1)^{w_1} f(S_1S_2)^{w_2} C(S_2|O_2)^{w_3} \quad (4)$$

Unlike the previous formula this score does not take into account the marginal probabilities. However these are indirectly used as the best results were obtained by setting the weight of the connection frequencies somewhat higher than the weights of the acoustic scores, but lower for very rare stations or stations having a high marginal probability.

7 Experiments

To concentrate on the problem at hand and not suffer from problems like locating the station names within an utterance we used simple recordings of station names for these experiments. In total we had 10163 recordings of station names. To get a representative dataset we selected a subset from the set of all possible combinations of names. This subset contains 325000 connections and exhibits roughly the same distribution as the OVR data. To test the stability of the methods another test set was selected, which has a much flatter distribution that the original dataset. Unlike the first set and the OVR data where 50% of all connections have a frequency above 1000, here only 26% of all connections have a high frequency. For the rescoring approaches the recognizer produced 10-best lattices.

Table 2. Comparing word error rates for the different approaches

System	WER Test set 1	WER Test set 2
Speech recognizer	11.4%	11.4%
Adaptive lm	10.34%	10.8%
Bi-directional lm	8.5%	9.9%
Highest confidence	10.9%	10.9%
Bayesian approach	9.5%	10.6%
Geometric mean	9.0%	10.4%

Table 2 shows the results of these experiments. The first row shows the word error rate of the baseline speech recognizer. The second row shows the word error rate if for each name the alternative with the highest confidence value is chosen. This already gives a small improvement as the word error rate is explicitly minimized this way. Including connection information using the Bayesian approach gives some improvements; the geometric mean does even better and results in an absolute gain of 2.5%. As can be expected both methods perform less well for the second data set, but they still improve upon the baseline system. The difference in performance between the two approaches stems from the ability to weight the contributions of knowledge

sources for the geometric mean. The mixture language model does less well than the post processing approaches, however when used for both station names it outperforms the other approaches. This is mainly due to the fact that in the post-processing approach the correct name may already be pruned away; in about 5% of all cases it is not in the n-best lists.

8 Conclusions

In this paper we described the use of station-to-station connection frequencies in a train table dialog system to explore how external knowledge sources can be used to enhance speech recognition. We experimented with a number of ways to incorporate such knowledge in the speech recognition process and to combine it with the acoustic likelihoods provided by the recognizer to come to a balanced decision. For this purpose a word level confidence measure was implemented.

The results suggest that an early integration approach for external knowledge offers the biggest potential. In our future research we want to continue in this direction, using knowledge from less restricted domains to come to a general model for utilizing contextual knowledge in speech recognition.

References

1. C. Chelba, F. Jelinek: Recognition Performance of a structured language model, Eurospeech '99, Budapest, Hungary
2. L. Mangu, E. Brill, A. Stolcke: Finding consensus among words, lattice-based word error minimization, In Proceedings of Eurospeech '99, Budapest, Hungary
3. G. Evermann, P.C. Woodland: Posterior probability decoding, confidence estimation and system combination, Proceedings of Speech Transcription Workshop, 2000
4. L.J.M. Rothkrantz, R.J. van Vark, A. Peters, A.C. Andeweg: Dialog control in the ALPARON System, Text Speech and Dialogs 2000
5. R.J. van Vark, J.P.M. de Vreught, L.J.M. Rothkrantz: Analysis of the OVR log files progress report, 1997, Alparon report nr. 97-07
6. P. Wiggers: A word level confidence measure for large vocabulary continuous speech recognition, 2003, Internal report DKS03-03
7. G. Tur, J. Wright, A. Gorin, G. Riccardi, D. Hakkani-Tür: Improving spoken language understanding using word confusion networks, In proceedings of ICSLP 2002
8. P.R. Clarkson, A.J. Robinson: Language model adaptation using mixtures and a exponentially decaying cache, In proceedings of ICASSP '97
9. G. Bouwman, L. Boves: Utterance verification Based on the Likelihood Distance to Alternative Paths, In proceedings of TSD 2002
10. G. Stemmer, S. Seidl, E. Noeth, H. Nieman, A. Batliner: Comparison and Combination of Confidence Measures, In proceedings of TSD 2002