

Effects of OOV rates on Keyphrase Rejection Schemes

Gies Bouwman, Janienke Sturm, Lou Boves

A²RT, Department of Language and Speech
University of Nijmegen, The Netherlands
{G.Bouwman, J.A.Sturm, L.Boves}@let.kun.nl
<http://lands.let.kun.nl/>

Abstract

Recognising directory listings for national telephone number inquiry is slowly getting within reach for modern ASR technology. Two key factors for a successful system design are (1) optimal extent of lexical modelling and (2) an effective utterance rejection method. In this paper we show how a choice for the first has consequences for the second.

We have taken the approach of building a lexicon with multiword expressions for the most frequently requested telephone listings, stepwise extended with filler words and less frequently addressed listings. In doing so, we keep track of the consequences that different Out of Vocabulary (OOV) rates have on two diverging keyphrase rejection schemes. Experimental results on field data clearly show that tasks with high OOV rates benefit most from acoustic confidence measures, while tasks with low OOV rates are better off with N-best list-based rejection schemes.

1. Introduction

In the framework of the EC-sponsored project SMADA (Speech-driven Multimodal Automatic Directory Assistance), we are investigating the feasibility of adopting ASR technology in a service for nation-wide Directory Assistance (DA). To make automation feasible, callers must first be prompted to say the type of listing (residential or business), next to give the city name and then to specify the name of a private person or a public agency, referred to as ‘residential listing’ and ‘business listings’ throughout this paper.

Automatically recognising names in DA is complicated in a number of ways.

1. the set of possible person or public agency names is very large,
2. the frequency distribution of requested listings is dependent on both location and time, and
3. especially for public agencies, there are often numerous possible expressions to refer to it.

When constructing a lexicon for a task like this, it is evident that optimal size is subject to a trade-off between coverage completeness and lexicon perplexity. The impact of the latter is amplified by the fact that callers often use spontaneous, brief, and ungrammatical sentences. In other words, it is difficult to capture the language in a model and thus to steer the process of selecting the right lexicon entry.

This consideration led us to the idea to build a lexicon with a small set of items we want to recognise, verify and extract information from. The lexicon covers the expressions (indicated as ‘keyphrases’ in the remainder of the paper) that refer to the X most frequently requested listings (FRLs). Han-

dling these listings automatically amounts to a large proportion of automation of the service. This lexicon can be extended with items we merely want to model. One could call the latter set ‘specified garbage’; it is just there to capture (some of) the incoming speech, but not to deduce information from.

Nevertheless, as a consequence of the problems mentioned above, many keyphrases are incorrectly recognised. Therefore, early rejection of keyphrases that may have been incorrectly recognised is an indispensable element of system design. In this paper we will examine the relation between the coverage of lexical modelling and the efficacy of different rejection schemes.

In order to reduce the complexity and the size of this problem, we will focus on a subset of our development material in this paper. Our corpus only contains listing requests where the caller answered the first question ‘For which city?’ with ‘Rotterdam’. Additionally, we only investigate the degree of coverage of the lexicon used by the recogniser; no attempts are made to use knowledge sources like the time of the day or the day of the week that the request was recorded.

The problem addressed in this paper can now be formulated as follows: How can the largest possible proportion of the top X FRLs be recognised with a confidence that is high enough for further automatic handling? We decompose this problem into two elementary questions:

- What is the effect of Out of Vocabulary (OOV) items on recognition performance?
- How much is the quality of different keyphrase rejection schemes influenced by the OOV rate?

The rest of this paper is organised as follows. In Section 2 we will discuss the two rejection schemes that will be used. Section 3 describes other features of our method. In Section 4 and 5 we will present the results and discussion. Finally, we summarise our method and draw conclusions in Section 6.

2. Keyphrase Rejection

The focus of the present paper is on the relation between the OOV rate of a recognition task and the suitability of keyphrase rejection (KR) methods. A suitable method rejects incorrectly recognised keyphrases as accurately as possible, i.e. without discarding too much of the correctly recognised data. We consider two methods. The first, to be referred to as **N-best based** [1][2][3], accepts a recognised keyphrase only if there is agreement upon its semantic information in a limited N-best list. The other, **LLR based** [4][5], computes phone and word confidence scores on the basis of frame-level Log Likelihood Ratios (LLRs) of the most likely sentence hypothesis, and accepts or rejects the keyphrases on the basis of the calculated confidence scores.

2.1. N-best based method

In this method, we assume that a word graph is generated during recognition. We further assume that for every arc (word) in the graph, there is an acoustic likelihood score available. We process the word graph with a keyphrase spotter. Depending on keyphrase content, the grammar of the keyphrase spotter returns a filler (empty value) or one or more semantically relevant attributes. In our case, an attribute is a unique representation of one of the top X FRLs. This helps to solve the problem that business listings can be referred to in different ways. Next, we extract an attribute-annotated sentence N-best list from the wordgraph, while preserving the according acoustic sentence likelihood scores. The decision to accept or reject the semantic result of the recognition depends on the degree of unanimity in the contents of the N-best list.

In this approach the restriction criteria applied when building the N-best list are a key element to set the balance of false accepts and false rejects. After all, the less restrictive criteria one uses, the more candidates in the list, the smaller the probability of agreement, the more false rejects and vice versa. The only parameter that we will investigate here to control the depth of the N-best list is the maximum distance to the acoustic likelihood score of the first best sentence that list members are allowed to have

2.2. LLR-based method

The other method, LLR-based, only considers the most likely sentence hypothesis, for which an ‘acoustic’ confidence score is computed for the keyphrase (if any) that it contains. We reject keyphrases with a confidence score lower than an a-priori set threshold value.

The score is obtained by using two acoustic models for each phone token in the keyphrase [4][5]. The first, to be indicated with ‘target model’, is equal or highly similar to the model used for recognition. The other, the ‘anti-model’, models all phones with which the phone under consideration is easily confused. For the feature vectors that are assigned to the phones of a keyphrase during recognition, we compute two likelihood scores with the corresponding target- and anti-models. The phone-level average of the ratio of these two frame-scores is also known as the Phone Likelihood Ratio. It represents the degree to which the hypothesised phone token is more likely to belong to the target than to its confusable competitors.

The Phone Likelihood Ratio score may be interpreted as a phone confidence measure (PCM). The PCMs of the N(W) phonemes of a word W are subsequently combined into a word confidence score C(W) according to formula (1).

$$C(W) = \frac{1}{\alpha} \ln \left(\frac{1}{N(W)} \cdot \sum_{i=1}^{N(W)} e^{\alpha \cdot PCM_i} \right) \quad (1)$$

With this weighting a minority of poor PCMs is enough to lower the whole word score. The constant α controls the sensitivity to local mismatches. In all of our experiments we used $\alpha = 0.4$. As a consequence, words that have been substituted with phonetically close (and thus confusable) words, may still be rejected for having local mismatches, as opposed to a simple average, where a couple of poor PCMs would not stand out sufficiently.

2.3. Comparison

The reason why we have chosen for the two KR methods described above, is that we consider them to be mutual extremes in a certain sense; the LLR-based method starts with phone units and likelihood distances at state-level. The other method takes semantic attributes with sentence-level scores as a starting point.

3. Method

3.1. Material

The material used for our experiments is a subset of the Dutch Directory Assistance Corpus (DDAC2000) [6]. Our research focuses on the part of the corpus that comprises the response to the prompt for the person or business name for all calls that pertained to the city of Rotterdam. In the present work, we will look at the top 190 FRLs for Rotterdam, that make up about 30% of all requests. One of the names is ‘unknown’; it covers for situations where a caller says that (s)he doesn’t know the name of the listing. The total corpus contains 3,121 utterances.

Recordings were made from the public switched telephone network. The signal was sampled at 8 kHz and stored in a-law format. Acoustic pre-processing comprised extracting 14 MFCCs (c0..c13) and their first-order derivatives from 16 ms Hamming windowed frames, with a 10 ms shift.

3.2. Acoustic and language models

Acoustic models were trained on 42,101 short utterances of the Dutch Polyphone database [6]. The HMM set consists of 37 tristate monophone models, one tristate noise and one single state silence model. In each state, acoustic variance is modelled by a mixture pdf of maximally 32 Gaussians. The decoder operated as a continuous speech recogniser and all lexicon items could be chosen equiprobably, i.e. we used a zero-gram language model for recognition.

3.3. Lexica

Since we want to investigate the suitability of KR methods for different OOV rates of the recognition task, lexicon coverage is the independent variable in our experiment. We decrease coverage by excluding a growing proportion of words used in the test utterances from the lexicon, making sure that the least frequently occurring word types are the first to be removed. However, in every lexicon we leave a set of 3 entries for filled pauses and noise and 891 (multi)word expressions intact. This set of expressions pertains to the top 190 FRLs. The majority of these expressions were created by hand on the basis of intuition. This fixed sublexicon covers only 26.2% of the 7,664 word tokens in the test corpus. The four experimental conditions are summarised in Table 1.

3.4. LLR models

This section describes how we defined and trained target and anti-models to compute the LLR based confidence scores.

For the anti-models, it is necessary to know with which phonemes each phoneme is typically confused. To determine the set of most confusable phones, the training corpus was segmented by the recogniser. Each phone was then scored against the models for all other phones. In this way, we ob-

Lexicon	Description	size	%OOV
rdam0%oov	top 190 FRL expressions completed with all other word types of the test corpus	3466	0%
rdam10%oov	top 190 FRL expressions completed with the most frequently occurring word types such that 90% of the word tokens are covered	2700	10%
rdam20%oov	top 190 FRL expressions completed with the most frequently occurring word types such that 80% of the word tokens are covered	1934	20%
rdam190	top 190 FRL expressions	894	73.8%

Table 1: Experimental lexica

tained a phoneme confusion list for each phone type. We used these lists in the following procedure.

We used the alignment described in the previous paragraph to compute a state-level likelihood score for each feature vector. Next, we computed phone scores by averaging frame scores at phone level. The tristate target models were subsequently trained on the best scoring 95% of all tokens, in order to reduce risk that mislabelled train tokens are included. During training, the state level segmentation of the material was kept unaltered. Then we trained single state anti-models on the best scoring 20% tokens of the 8 most confusable phoneme types. Target and anti-models are HMMs, each having mixture pdfs of maximally 32 Gaussians per state.

3.5. Evaluation metrics

We use error-based metrics to evaluate the performance of our recognisers. Keyphrase Error Rate (KER) and Word Error Rate (WER) will all be computed with formula (2):

$$\text{Error Rate} = (I + S + D) / N \times 100\% \quad (2)$$

where N is the total number of items, I the number of insertions, S substitutions and D deletions. Sentence Error Rate (SER) will also be evaluated.

For both rejection schemes, we use the same evaluation measure, viz. Rejection Error Rate (RER):

$$\text{RER}(T) = (D + \text{fr}(T) + \text{fa}(T)) / N \times 100\% \quad (3)$$

where T is a rejection threshold and N is the total number of items. $\text{fr}(T)$ is the number of falsely rejected items (a correctly recognised keyphrase or sentence was rejected) and $\text{fa}(T)$ is the number of falsely accepted items (an inserted or substituted keyphrase was accepted). We will compute minimal RER on keyphrase (RER_k) and on sentence (RER_s) level. A rejected sentence is considered falsely rejected when it contains at least one keyphrase. An accepted sentence is falsely accepted when all of the keyphrases are incorrect.

As pointed out in Section 2, the ‘threshold’ lies in different domains for the two keyphrase rejection schemes. In the LLR-based case, it is a real number, to which the outcome of (1) is compared. In the N-best case, it is a parameter that specifies the maximal likelihood score distance between the first best sentence and any other N-best list member. We will show a figure of RER_s plotted against several threshold values for both methods.

4. Results

This section presents the results of the experiments outlined in the previous sections. Table 2 shows the recognition performance for the best sentence. As can be seen, KER always exceeds 100%. The smaller the coverage of the lexicon, the more this can be ascribed to insertion errors.

Lexicon	SER	KER	WER
rdam0%oov	61.5%	131.2%	65.5%
rdam10%oov	66.0%	167.6%	71.4%
rdam20%oov	69.3%	205.5%	77.0%
rdam190	82.4%	380.7%	94.8%

Table 2: Recognition Error Rates (best sentence only)

Table 3 displays SER and WER for all experimental conditions when evaluating the whole wordgraph rather than just the best sentence. For now, we highlight just the fact that even when we use a lexicon that covers all word tokens of the test corpus, 30% of the correct words still do not appear in the wordgraph.

Lexicon	SER	WER
rdam0%oov	36.7%	30.0%
rdam10%oov	46.3%	37.2%
rdam20%oov	55.1%	45.0%
rdam190	79.6%	89.2%

Table 3: Recognition Error Rates (wordgraph)

Table 4 shows the rejection error rates after optimisation of the threshold.

Lexicon	LLR-based		Nbest-based	
	RER_s	RER_k	RER_s	RER_k
rdam0%oov	20.4%	72.6%	17.2%	63.0%
rdam10%oov	21.7%	77.5%	17.7%	65.6%
rdam20%oov	23.0%	81.7%	18.6%	68.7%
rdam190	25.7%	89.8%	27.2%	102.3%

Table 4: Rejection Error Rates

Finally, Figures 1 and 2 show RER_s as a function of the threshold value for the LLR-based and Nbest-based methods. We have no plots for RER_k , but we do remark that the curves of RER_s and RER_k are more or less parallel, having an optimal value for the same T in each of the four test conditions.

5. Discussion

The first thing to notice in Table 2 is the value of KER for lexicon rdam190. It can easily be explained: when trying to recognise key phrases where no different names were spoken, we end up with a keyword for almost every utterance. The number of insertion errors surpasses the number of actually spoken keywords, which leads to error rates $\gg 100$.

Another remarkable figure is the WER in the condition where none of the test utterances contain OOV items. Table 2 shows that very often the right word is not selected in the best

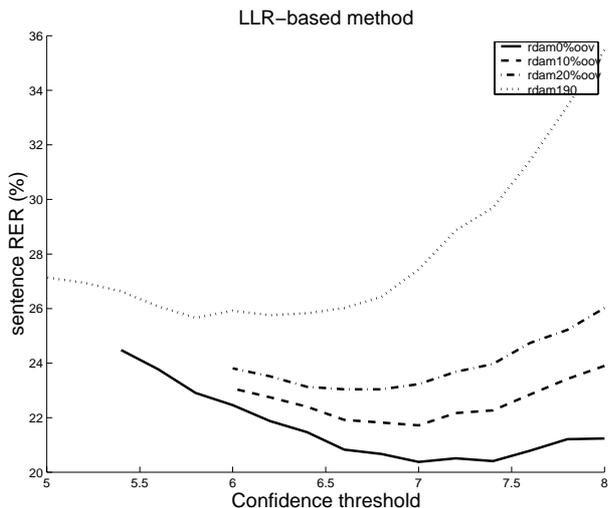


Figure 1: $RER_s(T)$ for the LLR-based measure

sentence, which may improve when using a more sophisticated language model. However, Table 3 shows that 30% of the spoken words do not appear in the word graph at all. In additional experiments the results of which are not reported in detail, we have seen that this is not caused by overrestrictive pruning criteria of the word graph generation, so we have no explanation for this phenomenon yet.

In [3] we hypothesised that the N-best method has a drawback that is compensated by the LLR-based method. It may happen that OOV utterances lead to word graphs with only one, obviously incorrect path. ‘Agreement’ in an N-best list with length 1 is automatically 100% and this inevitably leads to false accepts. Table 4 shows that the LLR-based rejection scheme performs better than the N-best method when the lexicon contains only the top 190 listing names. However, if the OOV rate is reduced by extending the lexicon, the performance of the N-best based method improves substantially. Our explanation is that the words added to the lexicon help to describe the formerly OOV speech better, or at least pose a threat to an erroneously hypothesised keyword. This decreases the number of false accepts much more than it increases the number of false rejects. A decrease in both RER_k and RER_s is the result.

Figure 2 shows that each RER stays the same beyond a certain threshold value. An explanation may be that all hypotheses in the original word graph have the common property that they are within the pruning criteria of the search beam used for recognition. It appears that this is a ‘hidden’ criterion for composing the N-best list as well. Beyond the concerned threshold value, the word graph has no more candidates that can cause rejection of the first best sentence. No more false rejects and false accepts are introduced, causing RER to stay the same. The results in Table 3 confirm this, as indicated by the limited number of correct sentences in the word graph.

So far even the best results in Table 4 are not good enough for operational DA services. However, there are several ways in which both recognition and rejection performance can be improved. Recognition performance can be improved by training context dependent models, instead of the context-independent models used in the experiments reported here. The performance of the LLR-based rejection scheme can be improved by computing acoustic confidence scores for all

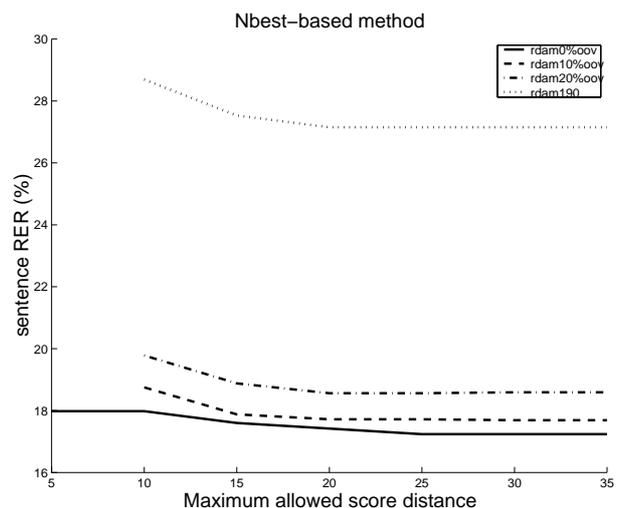


Figure 2: $RER_s(T)$ for the Nbest-based measure

entries of the N-best list, and reordering the list on the basis of the confidence measures.

6. Summary and conclusions

In this paper we have compared the performance of an LLR-based and an Nbest-based rejection scheme under four artificially created OOV rate conditions.

LLR-based measures are suitable for cases where an OOV item is recognised as an in-vocabulary word. Acoustic confidence measures will eliminate most of the ‘best yet wrong’ hypotheses. On the other hand, the N-best method has the advantage that it can handle acoustically similar hypotheses with incompatible semantic values. For the LLR-based measures, a confusable hypothesis may fit the speech signal so well, that it is not capable to ‘notice’ the substitution. Here is where the N-best method excels.

7. Acknowledgements

This research is supported by the EC under the IST-HLT Programme. Formula 1 is due to Chin Lee, of Lucent Bell Labs.

8. References

- [1] B. Rüber, *Obtaining Confidence Measures from Sentence Probabilities*, Proc. Eurospeech97, Rhodes, 1997, pp. 739-742
- [2] M. Weintraub, *LVCSR Log-Likelihood Ratio Scoring for Keyphrase Spotting*, Proc. ICASSP’95, Detroit, 1995, vol. I, pp. 297-300
- [3] G. Bouwman, J. Sturm, L. Boves, *Incorporating Confidence Measures in the Dutch Train Timetable Information System Developed in the Arise Project*, Proc. ICASSP’99, Phoenix, 1999, vol. I, pp. 493-496
- [4] Qi. Li et al., *Verbal Information Verification*, Proc. Eurospeech ‘97, Rhodes, Greece, vol. I, pp. 839-842
- [5] G. Bouwman, L. Boves, J. Koolwaaij, *Weighting Phone Confidence Measures for ASR*, Proc. COST249 Workshop on Voice Operated Telecom Services, Ghent, 2000, pp. 59-62
- [6] J. Sturm, H. Kamperman, L. Boves, E. den Os, *Impact of speaking style and speaking task on acoustic models*, Proc. ICSLP 2000, Beijing, 2000, pp. 361-364