TECHNIQUES FOR APPROXIMATING A TRIGRAM LANGUAGE MODEL

Fabio Brugnara and Marcello Federico

IRST – Istituto per la Ricerca Scientifica e Tecnologica
I–38050 Povo, Trento, Italy.

ABSTRACT

In this paper several methods are proposed for reducing the size of a trigram language model (LM), which is often the biggest data structure in a continuous speech recognizer, without affecting its performance. The common factor shared by the different approaches is to select only a subset of the available trigrams, trying to identify those trigrams that mostly contribute to the performance of the full trigram LM. The proposed selection criteria apply to trigram contexts, both of length one or two. These criteria rely on information theory concepts, the back-off probabilities estimated by the LM, or on a measure of the phonetic/linguistic uncertainty relative to a given context. Performance of the reduced trigrams LMs are compared both in terms of perplexity and recognition accuracy. Results show that all the considered methods perform better than the naive frequency shifting method. In fact, a 50% size reduction is obtained on a shift-1 trigram LM, at the cost of a 5% increase in word error rate. Moreover, the reduced LMs improve by around 15% the word error rate of a bigram LM of the same size.

1. INTRODUCTION

The language model (LM) is an important component of a speech recognition system. A-priori knowledge on the linguistic contents of incoming utterances can guide the search process that could not be focused enough if based on the acoustic evidence only. The most common way of modeling the stochastic process of sentence generation is by means of n-gram LMs, that have proven to be both powerful enough and easily integrable in the search process with the acoustic models. However, n-gram LMs have a number of free parameters that grows exponentially with n for a given dictionary size. This fact poses a severe limitation on the range of the order n, that is rarely bigger than 3. For dictation tasks, where the dictionary size often exceeds 10,000 words, even a trigram LM can have prohibitive size, due to the need of representing millions of different n-gram probabilities. In this paper, several methods are presented for approximating the modeling capabilities of a trigram LM by means of a reduced model, that involves only a fraction of the full set of probabilities. All methods rely on a selection of a suitable subset of the trigram probabilities, which are then used to augment a compact *shift-1* bigram LM, and differ in the criteria used in the subset selection.

2. LANGUAGE MODEL SIZE

In this paper, a standard interpolated trigram language model is considered, of the form:

$$\Pr(z \mid xy) = \begin{cases} fr^*(z \mid xy) + \lambda(xy)\Pr(z \mid y), & c(xy\_) > 0 \\ \Pr(z \mid y), & c(xy\_) = 0 \end{cases} \quad (1)$$

where $c(\cdot)$ denotes the absolute frequency of events, $fr^*(\cdot)$ denotes a discounted frequency, and $\lambda(\cdot)$ is the *back-off* probability. The counter $c(s\_)$ is used to specify all n-grams starting with a context s. Several methods exist for estimating LM probabilities, differing in the way the quantities fr^* and λ are computed. In this work, they are estimated with the standard *shift-1* method [3]:

$$fr^*(z \mid xy) = \frac{\max\{c(xyz) - 1, 0\}}{c(xy\_)} \quad (2)$$

$$\lambda(xy) = 1 - \sum_z fr^*(z \mid xy) \quad (3)$$

$$= \frac{\sum_{z:\, c(xyz)>0} 1}{c(xy\_)} \quad (4)$$

The bigram probability $\Pr(z \mid y)$ is computed similarly and employs for the back-off component a unigram probability estimated by adding 1 to all the single word counters.
Unigram, bigram and trigram statistics (counters) are stored separately, hence trigram selections do not have any effect on the lower order n-gram statistics. The following approximation methods select a subset of the trigram counters computed on the training set in order to reduce the size of the LM representation.
In many recognition systems, the LM (1) is represented through a *finite state network* whose number of transitions is dominated by the number of trigrams and bigrams for which the discounted frequency is greater than zero. This results from factorizing the interpolation formula (1) in terms of the back-off and the lower order probabilities [4]. In the following,

the number of bigrams and trigrams explicitly represented by the LM will be used to measure the size of the LM.

2.1. Frequency Shifting

The simplest method to reduce a trigram LM is shifting down all the trigram counters by a certain value. For instance, the discounting method (3) already shifts down all counters by 1. The rationale of this method is that LM performance should not be affected by eliminating rare events from the training sample. Moreover, this method provides large reductions as rare events constitute the majority of events. In order to establish a baseline for comparison with the following selection methods, *shift-2* and *shift-3* LMs were computed. The sizes of the resulting trigram sets were then assumed as references for choosing the selection thresholds.

3. TRIGRAM SELECTION

While the *shift-k* method achieves a reduction of the trigram set by modifying the counters, the following methods aim at identifying a set of contexts for which the insertion of trigram probabilities is mostly useful by computing, with different criteria, a scoring function $F(s)$ for every context s. The contexts are then sorted in order of decreasing scores, and trigrams are inserted by adding the extensions of the first contexts in the list, as to reach the desired LM size. Contexts s can be of type $\_y$ or xy. In order to avoid the selection of too many rare events, only trigrams such that $c(xy\_) > 10$ have been considered for selection.

4. FREQUENCY BASED SELECTION

The simplest method directly uses absolute probabilities as scores, and can be applied to both types of contexts, leading to the following scoring functions:

$$\begin{aligned} F_{F1}(\_y) &= \Pr(y) \\ F_{F2}(xy) &= \Pr(xy) \end{aligned}$$

In this way, no information about the conditional probability distribution function (pdf) $\Pr(\cdot \mid s)$ is taken into account.

5. BACK-OFF BASED SELECTION

A measure of the weakness of bigram probabilities given a context $\_y$ is the probability of resorting to unigrams in evaluating LM scores, that is expressed by the score function:

$$F_{B1}(\_y) = \Pr(y)\lambda(y) \tag{5}$$

According to the employed discounting method (3), the result is proportional to the number of different successors that have been observed after the context y.

6. INFORMATION BASED SELECTION

6.1. Entropy

A way of evaluating the degree of ambiguity of a context is to measure the entropy of the current word $\mathbf{w}_t$ given that context. The entropy of a bigram LM is:

$$H(\mathbf{w}_t \mid \mathbf{w}_{t-1}) = \sum_{w_{t-1}} \Pr(w_{t-1}) H(\mathbf{w}_t \mid w_{t-1}) \tag{6}$$

where

$$H(\mathbf{w}_t \mid w_{t-1}) = -\sum_{w_t} \Pr(w_t \mid w_{t-1}) \log \Pr(w_t \mid w_{t-1}) \tag{7}$$

A context of type $\_y$ can be considered the more ambiguous the highest its contribution to the global entropy is, suggesting the use of the following quantity as a scoring function:

$$F_{H1}(\_y) = \Pr(y) H(\mathbf{w}_t | \mathbf{w}_{t-1} = y) \tag{8}$$

The rationale of this criterion is that trigrams are mostly useful when bigrams provide high entropy or, similarly, high perplexity.

6.2. Mutual Information

The utility of introducing trigrams in a LM can also be measured with the amount of information about the current word ($\mathbf{w}_t$) gained by augmenting the context from ($\mathbf{w}_{t-1}$) to $\mathbf{w}_{t-2}\mathbf{w}_{t-1}$. This mutual information can be expressed as:

$$\begin{aligned} &I(\mathbf{w}_t; \mathbf{w}_{t-2} \mid \mathbf{w}_{t-1}) = \\ &= \sum_{w_{t-1}} \Pr(w_{t-1}) I(\mathbf{w}_t; \mathbf{w}_{t-2} \mid w_{t-1}) && (9) \\ &= \sum_{w_{t-2}^{t-1}} \Pr(w_{t-2}^{t-1}) I(\mathbf{w}_t; w_{t-2} \mid w_{t-1}) && (10) \\ &= \sum_{w_{t-2}^{t-1}} \Pr(w_{t-2}^{t-1}) (H(\mathbf{w}_t \mid w_{t-1}) - H(\mathbf{w}_t \mid w_{t-2}^{t-1})) \end{aligned}$$

Equations (9) and (10) explain the individual contribution to the global information of contexts of type $\_y$ and xy, respectively. The following score functions can thus be considered:

$$\begin{aligned} F_{I1}(\_y) &= \Pr(y) I(\mathbf{w}_t; \mathbf{w}_{t-2} \mid \mathbf{w}_{t-1} = y) \\ F_{I2}(xy) &= \Pr(xy) I(\mathbf{w}_t; \mathbf{w}_{t-2} = x \mid \mathbf{w}_{t-1} = y) \end{aligned}$$

7. PHONETIC SIMILARITY BASED SELECTION

The above methods only take into account, in different ways, the features of the LM pdf. However, the chance of recognition errors also depends on the acoustic similarity of words. A method is proposed that tries to consider both factors. A

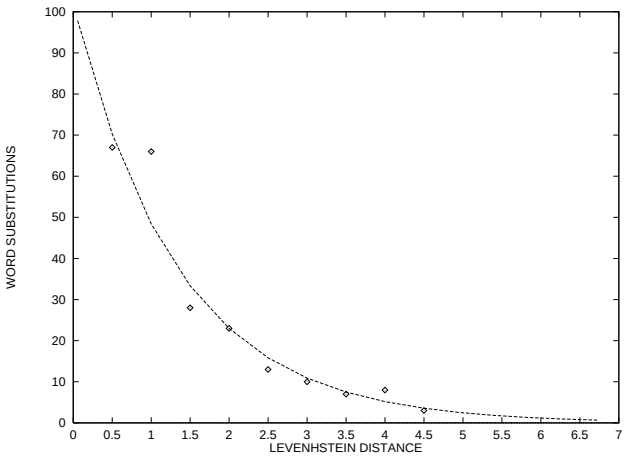

Figure 1: Substitution errors made by an isolated word recognizer versus Levenshtein distances of the phonetic transcriptions. The plotted curve represents the best fitting exponential regression curve.

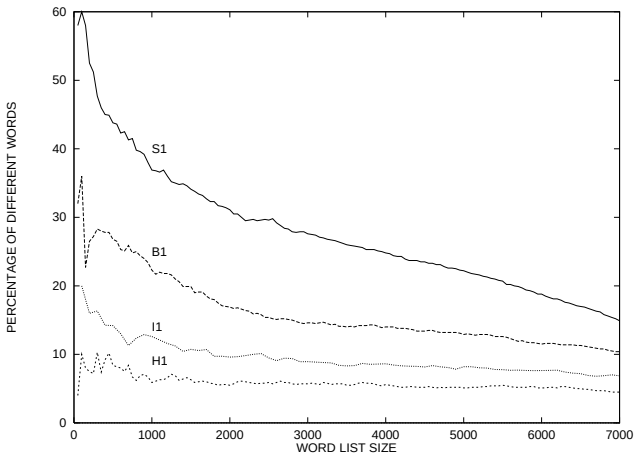

Figure 2: Differences between the criteria extending single contexts _y. The word ordering of each scoring function is compared with the frequency criterion (F1) for increasing list sizes.

phonetic/linguistic uncertainty measure for a word context y has been defined as follows:

$$S(y) = \sum_{z} \Pr(z \mid y) \max_{w \neq z} \Pr(w \mid y) s(w, z) \qquad (11)$$

and the relative scoring function is:

$$F_{S1}(\_y) = \Pr(y) S(y) \qquad (12)$$

The term $s(w, z)$ denotes a phonetic similarity measures between w and z, computed as:

$$s(w, z) = e^{-\alpha l(w,z)} \qquad (13)$$

with $\alpha > 0$ and $l(w, z)$ representing the Levenshtein distance between the phonetic transcriptions of w and z. The Levenshtein distance between two phonetic transcriptions is computed by assuming a penalty 1 for deletions and insertions, a penalty 0.5 for substitutions within phonemic classes - i.e. vowels, nasals, liquids, fricatives, voiced stops, and unvoiced stops -, and a penalty 1 for other substitutions. In order to define a similarity measure, a negative exponentiation has been chosen. An empirical evidence of this function has been found by looking at the recognition errors made by an isolated word speech recognizer working without LM. By grouping together all the different word substitution errors corresponding to a fixed Levenshtein distance, the values plotted in Figure 1 were obtained. Regression analysis shows that such points are well fitted by an exponential curve (see Figure 1).

In order to balance the influence of linguistic and phonetic scores in the computation of $F_{S1}(y)$, a decay constant α larger than that of the fitting curve was selected, which let $s(w, z)$ go very close to zero for distance values larger than 4.5. For practical reasons, only Levenshtein distances less than or equal to 3.5 were precomputed and stored for all possible word pairs. Moreover, in order to speed up computations, approximated values of $F_{S1}(y)$ were computed, providing relative errors less that 0.01.

8. EXPERIMENTS

All selection criteria have been empirically evaluated on a 10,000-word Italian newspaper dictation task [1]. Trigram selection methods have been applied for generating LMs of different sizes. LMs have been estimated on a 25 million word corpus.

As reference points, the trigram LM in (1) and a full bigram LM have been considered. The full bigram LM has been estimated with the non linear discounting formula proposed by Ney *et al.* [3]. A constant $0 < \beta < 1$ is subtracted from all bigram countings as (3) does, but, as a difference, no bigram counters are completely deleted. In fact, previous work showed that this non linear discounting approach provided the best performance over a range of different bigram LMs [2].

8.1. Qualitative Evaluation

A qualitative evaluation of the trigrams selection methods has been carried out by comparing the context orderings provided by each criterion. In Figure 2, the back-off (B1), the mutual information (I1), the entropy (H1), and the phonetic similarity (S1) based scoring functions are compared versus the frequency based (F1) one. It can be seen that the information based scores result to be the most influenced by the frequency of words.

8.2. Perplexity Tests

Perplexity evaluations were carried out on a testing set of about 1.6 million words. The different methods provided

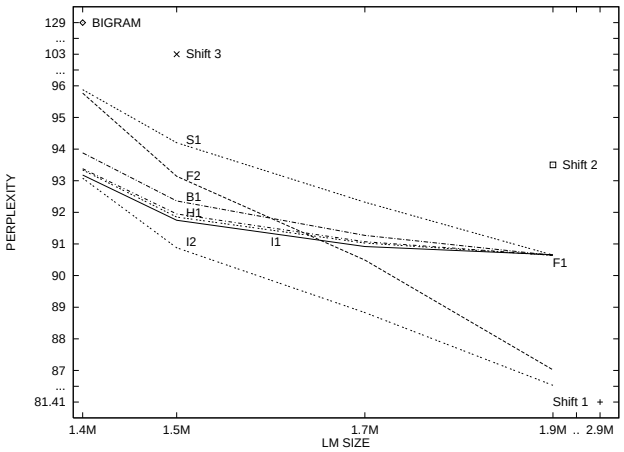

Figure 3: Perplexity evaluation of each approximation method with respect to the LM size (million of 2-grams and 3-grams).

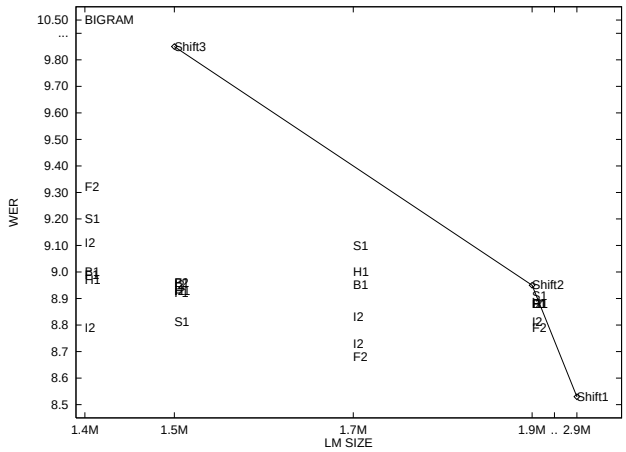

Figure 4: Word error rate of each approximation method with respect to the LM size.

perplexity values that are plotted in Figure 3. For fixed LM sizes all the trigram selection criteria perform better than the Shift-k method (for k=2,3) and than the full bigram LM. Moreover, the I2 criterion appears to provide the best trigram selection method. In fact, the F1, H1, and I1 criteria give very similar results, while the F2 criterion performs well only for large LM sizes. The S1 method performs quite worse, but this was expected as its trigram selection criterion is less related to perplexity and likelihood than its competitors.

8.3. Recognition Tests

The same LMs were compared in terms of recognition accuracy by employing them in a speaker independent, continuous speech recognizer [1]. In this case, the size of the testing material is not comparable with that used to measure perplexity. It consists of 360 sentences, uttered by 12 speakers in an office environment, for a total amount of $1^h : 15'$ of speech. A diagram of the measured word error rates (WER) is shown in Figure 4. The main result is that all the proposed methods allow to considerably reduce the LM size of the baseline trigram LM without considerably affecting the recognition accuracy. Moreover, all methods outperform the frequency shifting methods, and also show much better performance than a bigram LM with comparable size. However, the experiments do not allow to assess significant differences between the selection criteria. None of the methods performs uniformly better than the others, and not always larger LMs selected with the same method provide better performance, as on the contrary happens with perplexity. This is probably due to the limited size of the employed test set. The phonetic similarity based selection (S1) performs close to the other criteria in spite of being based on a quite different approach. In fact, as can be seen from Figure 2 the list of selected trigrams differs significantly from the other methods. It is to be noted, however, that this approach generates LMs with higher perplexity, as shown if Figure 3. The fact that the recognition results are nevertheless average suggests that the introduction of the phonetic similarity in the scoring function can compensate for the loss in linguistic focusing. Hence, it is possible that a more suitable choice of substitution penalties between acoustic units give better results.

9. CONCLUSIONS

Different criteria have been proposed to approximate a trigram LM by selecting the best trigram statistics to be represented explicitly. The deriving methods appear to be a better alternative to the naive frequency shifting method. In fact, trigrams LMs of size comparable to a bigram LM can be generated that perform much better, achieving a 15% WER reduction. Moreover, a LM size reduction of more than 50% can be achieved with respect to a Shift 1 trigrams LM, at the cost of a 5% loss in WER.

10. REFERENCES

1. G. Antoniol, F. Brugnara, M. Cettolo, and M. Federico. Language model representations for beam-search decoding. In *Proc. of the ICASSP*, pages I:588–591, Detroit, MI, May 1995.

2. M. Federico, M. Cettolo, F. Brugnara, and G. Antoniol. Language modeling for efficient beam-search. *Computer Speech and Language*, 9:353–379, 1995.

3. H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8:1–38, 1994.

4. P. Placeway, R. Schwartz, P. Fung, and L. Nguyen. The estimation of powerful language models from small and large corpora. In *Proc. of the ICASSP*, volume II, pages 33–36, Minneapolis, MN, 1993.