Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition Letters 25 (2004) 15–19

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Data augmentation and language model adaptation using singular value decomposition

F. Béchet, R. De Mori *, D. Janiszek

LIA-CNRS, University of Avignon, BP 1228, 84911 Avignon Cedex 9, France

Received 13 March 2003; received in revised form 5 August 2003

Abstract

A new augmentation method for counts to be used in language modeling is presented. It is based on word representations in a reduced space obtained with Singular Value Decomposition. A contribution to a count for a linguistic event x is obtained from the counts of observed events smoothed with a function of their distance from x. Experimental results on a spoken dialogue corpus show the performance of the proposed method, combined with maximum a posteriori probability adaptation, in terms of word error rate reduction.

Keywords: Statistical language modeling; Singular value decomposition; Data augmentation; Automatic speech recognition

1. Introduction

Most of the existing automatic speech recognition (ASR) systems generate word hypotheses with a Language Model (LM) which computes the probability of a sequence of words $W_1^N = w_1, w_2, \dots, w_n, \dots, w_N$ as follows:

$$P(W_1^N) = P(w_1)\prod_{n=2}^{N} P(w_n \mid w_1, w_2, \dots, w_{n-1}) \qquad (1)$$

where the sequence $w_1, w_2, \dots, w_{n-1}$ is called the *history* h_n of word w_n. A word can have many histories, thus the generic jth history of word w_n will be indicated as h_{nj}.

Usually, the probabilities appearing on the right-hand side of (1) are estimated with a training corpus. For long histories, it is practically impossible to find enough data in an even very large corpus, and approximations are introduced by clustering all the histories having the same one or two last words resulting in well-known bigram or trigram LMs.

When a new application domain is considered, many bigram and trigram probabilities change and new corpora are required for obtaining appropriate LMs. If large corpora are not available for this purpose, then an LM can be obtained by adapting available LMs trained with a large corpus in a more general domain. Various methods for LM adaptation have been proposed and an overview can be found in (Bellegarda, 2001). In general,

* Corresponding author. Address: LIA-CERI, University of Avignon, 339 Chemin des Meinajaries, BP 1228, 84911 Avignon Cedex 9, France. Tel.: +33-4-9084-3515; fax: +33-4-9084-3501.

E-mail address: renato.demori@lia.univ-avignon.fr (R. De Mori).

0167-8655/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2003.08.008

even if a large training corpus is available, a number of linguistic events modeled by the LM are likely to be absent in the corpus.

Instead of adapting LM parameters, it is possible to perform *data augmentation* by inferring counts for training, based on the available adaptation data, in such a way that LM probabilities are estimated from counts obtained only from the adaptation data augmented with counts generated by a suitable *smoothing/generalization* criterion. Data augmentation is intended here as the computation of counts for unseen linguistic events using available counts of events that have been observed in a limited, application dependent corpus.

The approach proposed in this letter is based on the conjecture that if a word has been observed in a given context, then semantically similar words are likely to appear in the same context even if this event was not observed in the adaptation corpus.

Semantic similarity between words can be defined using a numerical distance between vectors representing words in a suitable space. Following an approach of Information Retrieval (Berry, 1992; Bellegarda, 1998; Deerwester et al., 1990), such a space can be defined using *Singular Value Decomposition* (SVD). In this way, the counts of the general purpose corpus and the counts obtained with adaptation have the same representation.

Further improvements can be obtained by performing maximum a posteriori (MAP) probability adaptation on the LM obtained with data augmentation.

2. Data augmentation

Let $\boldsymbol{P} = \{p_{ij}\}$ be a $I \times J$ matrix where the generic element $\{p_{ij}\}$ represents the probability or, simply, the count of observations of word w_i in the context of history h_j. Empirical evidence has shown that using counts provided better results than using probabilities indicating that normalization introduced by probability computation from counts is not effective. Thus, $\boldsymbol{P}$ was built as a matrix of counts. The ith row of matrix $\boldsymbol{P}$ is a vector whose J elements are the probabilities or the counts of w_i with all possible histories.

By considering only the q prominent eigenvalues of $\boldsymbol{P}$, a diagonal matrix $\boldsymbol{S}$ can be built with the first q singular values in decreasing order such that $\boldsymbol{P} \cong \boldsymbol{USV}^{\mathrm{T}}$ where the $\boldsymbol{U}$ has q columns consisting of the first q eigenvectors while $\boldsymbol{V}$ is made with the first q eigenvectors.

Matrices $\boldsymbol{U}$, $\boldsymbol{S}$ and $\boldsymbol{V}$ are computed with an iterative procedure as proposed in (Berry, 1992) for a value of q chosen in a such a way that s_0/s_q is (s_q being the qth singular value) approximately equal to 10^3. This value has been found to be a reasonable compromise between accuracy and computational complexity.

Vector $\boldsymbol{P}_i$ can be represented in reduced space by vector $\boldsymbol{R}_i$ obtained as follows: $\boldsymbol{R}_i = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{P}_i$. Because the number of columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ is much smaller than the number of columns in $\boldsymbol{P}$, it is expected that not many elements of $\boldsymbol{R}_i$ are equal to zero.

If $\boldsymbol{P}$ has been trained on a very large corpus containing a good mix of topics, one may assume that the estimated eigenvalues which are the non-zero elements of matrix $\boldsymbol{S}$, are typical of a language and do not vary from one application to another. This conjecture has been validated experimentally using a corpus made of 40 million words from articles of the French newspaper *Le Monde* (60 kword lexicon) and corpora made of telephone utterances from person–machine dialogs collected in fairly severe conditions all over France.

Let c_{ij} be the count of times the word w_i has been really observed in the adaptation data in the context of history h_{ij}. Let a_{ij} be the count for the same word and history, but after data augmentation. Let $\Gamma_j^{\alpha}(\vartheta)$ be the set of vectors representing the histories whose distance from the vector representing h_{ij} in the reduced space satisfy a property ϑ. Let d_{jk}^{α} be the distance between vectors representing histories h_{ij} and h_{ik} in reduced space α.

If the LM contains only bigrams, then $h_{ij} = w_j$. The augmented count a_{ij} of the sequence $[w_j w_i]$ is obtained assuming that a history w_k *similar* to w_j contributes to the counts of the sequence $[w_j w_i]$ in a way that depends on a degree of similarity between the two histories w_j and w_k:

$$a_{ij} = c_{ij} + \sum_{h_{ik} \in \Gamma_j^{\alpha}(\vartheta)} c_{ik} \cdot f(d_{jk}^{\alpha}) \tag{2}$$

Such a degree of similarity is represented by a function $f(d^{\alpha}_{jk})$ of the distance between the representations of the two histories. The function $f(d^{\alpha}_{jk})$ should be equal to 1 when $d^{\alpha}_{jk} = 0$ and should decrease with d^{α}_{jk}. A reasonable assumption is the following:

$$f(d^{\alpha}_{jk}) = \mathrm{e}^{-d^{\alpha}_{jk}/D} \tag{3}$$

where D is a decay that can be used for tuning the system.

The use of an exponential function was inspired by past experience on softmax functions applied at the output layer of neural networks and has the same motivations.

It was observed that distances of words to a given word do not have a uniform distribution. There is often a cluster of close words which are likely to be semantically similar. It was observed that only the contribution of these words may lead to improvements with little differences if D varies between 0.8 and 1.2. This also suggests that the choice of an exponential function is not critical.

The K selected contributions correspond to the above mentioned cluster. K may vary from word to word.

The Euclidian distance between each pair of history vectors was computed in reduced space. The angle between each pair of history vectors was also considered because it is used in Information Retrieval but it was abandoned because it produced poor results.

An ASR system called AGS and developed at France-Telecom R&D was used for generating a trellis of word hypotheses as well as the best hypothesis for each spoken sentence. Let such a baseline system be indicated as B. Let $T_B(k)$ be the trellis provided by the base system for the kth sentence. The baseline system uses a bigram LM obtained with the entire training corpus.

The new LMs obtained after data augmentation was used for searching the best path in $T_B(k)$ by using the same acoustic scores provided when $T_B(k)$ was generated.

Sentences of the test set could be divided into two groups. The fist one contains all the sentences for which the correct transcription is a sequence of words that corresponds to an existing path in the trellis. The second groups contains sentences for which such a path does not exist. As the sentences have been acquired on the field in the French network, many sentences containing errors are in the second group because one or more words were truncated, as a consequence of false (late) starts or premature termination of the decoding process. Obviously, there is no way to recover such errors with a trellis rescoring process.

The training set used for estimating the LM parameters consists of 9842 sentences for a total of about 70,000 words. Tests were performed on 229 telephone sentences, for a total of 2031 words, having a corresponding path in the trellises used for rescoring.

A first experiment was performed with $D = 1$ and contribution from all histories. Results did not show a strong improvement with data augmentation, but suggested that history dependent thresholds should be used.

Interesting results were obtained by augmenting each bigram count with contributions from the K *nearest histories*. As it may happen that, for a given K, there are many histories with very close distance w.r.t. the Kth one, contributions from all these histories were also considered after having empirically selected a threshold for considering distances to be practically equivalent. Let us call this approach *quasi-K nearest histories*.

Experiments were conducted using LMs built by randomly picking a portion of the training set containing 70,000 words. For each set, the LM was adapted with data augmentation. Fig. 1 reports the Word Error Rate (WER) as function of the training set size (in thousands of words) indicated as *set size*, with and without data augmentation. WER is defined as follows:

$$\mathrm{WER} = \frac{N_{\mathrm{S}} + N_{\mathrm{D}} + N_{\mathrm{I}}}{N_{\mathrm{W}}} 100$$

where N_{W} is the total number of words in the test set, N_{S} is the total number of word substitutions observed after running a recognition experiment on the test set, N_{D} is the total number of word deletions observed after running a recognition experiment on the test set, N_{I} is the total number of word insertions observed after running a recognition experiment on the test set.

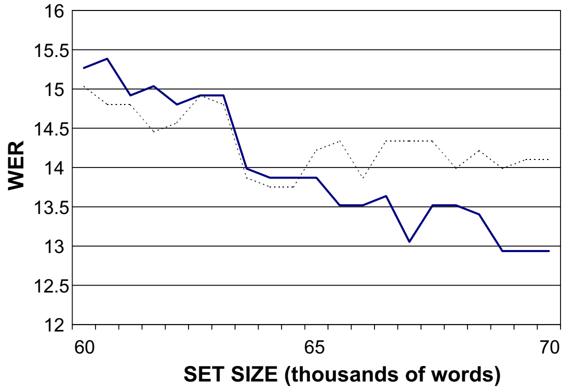

Fig. 1. Word Error Rate as a function of the size (in thousands of words) of the set used for training the LM. The dotted line represents the results with the LM probabilities estimated using only the application dependent corpus without data augmentation. The thick line represents the results of data augmentation using *quasi-K nearest histories adaptation*.

Experimental results show that working in a reduced space derived only with application data does not result in a tangible improvement suggesting that the reduced space obtained with *Le Monde* data is good enough for other types of domains. Data augmentation benefits become tangible when the training set size is larger than 65,000 words.

With augmented counts, some of the constraints that should hold between row and column marginal counts may no longer hold. Even if in practice, the discrepancy is small between the sum of counts for a word row and the sum of counts in the column corresponding to the same word considered as history, it is possible to reestablish constraint satisfaction. Details are omitted here for the sake of brevity.

3. Adaptation with maximum a posteriori probability

In (De Mori and Federico, 1999), it is shown that MAP adaptation of LM probabilities can be performed by a linear interpolation of the a priori probabilities provided by the general LM and the probabilities obtained with the adaptation corpus. The same idea can be applied to bigram counts. Let $c_{\mathrm{g}}(w_j, w_i)$ and $c_{\mathrm{d}}(w_j, w_i)$ be respectively the bigram counts in the general corpus and in the domain adaptation corpus. Let N_{g} and N_{d} be respectively the sizes of the general corpus and the domain adaptation corpus.

MAP adaptation can be performed by interpolation of the general model counts and the counts of the adaptation corpus, leading to the following expressions for adapted counts:

$$P_{\mathrm{MAP}}(w_i|w_j) = \frac{N_{\mathrm{g}}}{N_{\mathrm{g}}+N_{\mathrm{d}}}\frac{c_{\mathrm{g}}(w_j,w_i)}{c_{\mathrm{g}}(w_j)} + \frac{N_{\mathrm{d}}}{N_{\mathrm{g}}+N_{\mathrm{d}}}\frac{c_{\mathrm{d}}(w_j,w_i)}{c_{\mathrm{d}}(w_j)} \tag{4}$$

No improvements were observed by applying this general model for adapting a LM trained with the corpus from *Le Monde* (g) using the training corpus of domain specific (d) data for adaptation.

Nevertheless, another 10% WER reduction was obtained by using counts from some specific histories of ‘g’ to augment counts for the same histories of ‘d’. In fact, (4) can be applied separately for each history and can be re-written as follows:

$$c_{\mathrm{a}}(w_i, w_j) = \zeta(w_j)c_{\mathrm{g}}(w_j, w_i) + \{1 - \zeta(w_j)\}c_{\mathrm{d}}(w_j, w_i) \tag{5}$$

where $\zeta(w_j)$ is a linear combination weight determined with maximum likelihood estimation on a development set. If such a set is too small as in our case, then a common $\zeta(w_j)$ is estimated for all words.

Pertinent histories were obtained with a small development set. It is important to point out that all the methods presented in this letter do not lead to any increase in the memory space required for storing the LM probabilities.

4. Conclusions

A simple method has been proposed for obtaining bigram counts of unseen linguistic events by inference from counts of observed events weighted with a function of similarity between words. When counts are used for estimating the probabilities of a bigram LM, this approach provides estimates based on counts which avoid the use of back-off. Experimental results support the advantage of this approach showing a tangible WER reduction. Further benefits can be obtained by performing MAP adaptation of a general purpose LM using augmented counts inferred from a domain specific corpus.

LMs built in this way have proven to be useful also for performing sentence verification based on the consensus of recognition results obtained with different LMs (Esteve et al., 2003).

References

Berry, M.W., 1992. Large-scale sparse singular value computations. Int. J. Supercomput. Appl. 6 (1), 13–49.

Bellegarda, J., 1998. Multi-span statistical language modeling for large vocabulary speech recognition. IEEE Trans. Speech Audio Process. SAP-6 (5), 456–467.

Bellegarda, J., 2001. An overview of statistical language model adaptation. In: Proc. ISCA-ITR Workshop on Adaptation Methods for Speech Recognition, Sophia-Antipolis, France, August, pp. 165–175.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by Latent Semantic Analysis. J. Am. Soc. Inform. Sci. 41, 391–407.

De Mori, R., Federico, M., 1999. Language model adaptation. In: Ponting, K. (Ed.), Computational Models of Speech Pattern Processing. Springer-Verlag, Berlin, New York, p. 1999.

Esteve, Y., Raymond, C., De Mori, R., Janiszek, D., 2003. On the use of linguistic consistency in systems for human–computer dialogs. IEEE Trans. Speech Audio Process, in press.