DATA AUGMENTATION AND LANGUAGE MODEL ADAPTATION

D. Janiszek, R. De Mori, F. Bechet

LIA - University of Avignon
84911 Avignon Cedex 9 - France

ABSTRACT

A method is presented for augmenting word n-gram counts in a matrix which represents a 2-gram Language Model (LM). This method is based on numerical distances in a reduced space obtained by Singular Value Decomposition (SVD). Rescoring word lattices in a spoken dialogue application using an LM containing augmented counts has lead to a Word Error Rate (WER) reduction of 6.5%. By further interpolating augmented counts with the counts extracted from a very large newspaper corpus, but only for selected histories, a total WER reduction of 11.7% was obtained. We show that this approach gives better results than a global count interpolation for all histories of the LM.

1. INTRODUCTION

Most of the existing automatic speech recognition (ASR) systems generate word hypotheses by computing the probability of a sequence of words W_1^N as a product of conditional probabilities of words w_i and their histories h_i. Language models (*LM*) provide probability distributions $P(w_i|h_i)$ for each word w_i of the vocabulary and for each history class h_i. Probabilities $P(w_i|h_i)$ are estimated using a training corpus. In practice, their value and the accuracy of the estimation depend on the corpus.

Usually an LM trained in a domain D_1 exhibits poorer performance when applied to a domain D_2 with respect to an LM trained in the new domain D_2. This is also true in dialogue systems in which the probability of words and histories often depend on dialogue situations, especially when the machine asks questions about specific topics. Different solutions exist to overcome the problem of LM variations. If large corpora are not available, then LMs can be adapted. Various methods for LM adaptation have been proposed and reviews of them can be found in [1]. Most of the proposed methods perform dynamic adaptation, consisting in continuously updating in time the LM probability distributions.

There are many methods for adapting to a new domain which can be grouped into the following classes:

This research is supported by France Telecom's R&D under the contract 971B427

- training an LM in the new domain if sufficient data are available,
- pooling data of many domains with data of the new domain,
- linear interpolation of a general and a domain-specific model [2],
- back-off of domain specific probabilities with those of a general model [3],
- retrieval of documents pertinent to the new domain and training a new LM on-line with those data [4],
- maximum entropy, minimum discrimination adaptation [5],
- Maximum A Posteriori Probability (MAP) adaptation [1],
- adaptation by linear transformation of vectors of bigram counts in a reduced space [6].

Another possibility could be that of performing *data augmentation* by inferring counts for the training set 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. The approach proposed in this paper starts with 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 [7], such a space can be defined using *Singular Value Decomposition* (SVD).

Given a vector representing a word W, it is possible to define a cone around it and consider all words represented by vectors inside this cone as semantically similar to W. The similarity between a word $W^{'}$ in the cone and W can be expressed by a distance between the vectors representing the two words. The count of W in a given context can thus

be augmented by a contribution from the count of $W^{'}$ in the same context weighted by a decreasing function of the distance between the vectors representing the two words. Section 4 details this approach.

It is important to point out that all the methods presented in this paper do not lead to any increase in the LM complexity.

2. REPRESENTATION OF WORDS IN REDUCED SPACE

Representation of words by vectors in a reduced space can be obtained [8], [7] by SVD based on which any matrix P of dimensions I, J and rank r can be expressed as:

$$P = L \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} M^T \quad (1)$$

where L and M are orthogonal matrices, their columns are orthonormal, M^T indicates the transpose matrix of M, and Δ is the diagonal matrix of the positive eigenvalues of PP^T. The order of Δ is $\{r, r\}$. The columns of L are an orthonormal set of basis vectors spanning the range. By considering only the q prominent eigenvalues, a diagonal matrix S can be built with the first q singular values of PP^T in decreasing order such that:

$$P \cong USV^T \quad (2)$$

where the U has q columns consisting of the first q eigenvectors of PP^T while V is made with the first q eigenvectors of P^TP.

Matrices U, S and V are computed with an iterative procedure as proposed in [8] for a value of q chosen in a such a way that s_0/s_q (with s_q the q-th singular value) approximate 10^3, which has been found to be a reasonable compromise between accuracy and computational complexity.

Matrix P is, in our case, a matrix having rows corresponding to words and columns to histories. It can be used to find the transformation for mapping word and history vectors into the reduced space.

Let $P = \{p_{ij}\}$ be a $I \times J$ matrix where the generic element $\{p_{ij}\}$ represents the count of observations of word W_i in the context of history h_j. The i-th row of matrix P is a vector whose J elements are the counts of w_i in all possible histories.

Vector P_i can be represented in reduced space by vector R_i obtained as follows:

$$R_i = U^T P_i \quad (3)$$

Because the number of columns of U and V is much smaller than the number of columns in P, it is expected that not many elements of R_i are equal to zero.

If 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 S, are typical of a language and do not vary from one application to another. This conjecture has been validated experimentally as it will be shown later on using a corpus made of 40 million words from articles of the French newspaper *Le Monde*.

3. EXPERIMENTAL SETUP

A system, called *AGS*, described in [9], and deployed on the telephone network, performed a fist step recognition for the test set, and made available, for each test sentence, a trellis of word hypotheses as well as the best hypothesis produced by the system. Let such a baseline system be indicated as B. Let $T_B(k)$ be the trellis provided by the base system for the k-th sentence. The purpose of the experiments is that of assessing if and how much the Word Error Rate (WER) can be reduced by rescoring the word hypotheses in $T_B(k)$ with new LMs obtained after data augmentation or adaptation, by using the same scores provided by the acoustic models when $T_B(k)$ was generated.

Experiments were carried out using a telephone corpus of sentences from person-machine dialogues collected by France-Telecom R&D in fairly severe conditions all over France.

The corpus contains a training set of 9842 sentences. Another set of 1419 word lattices produced by the AGS system was arbitrarily split in a 1000 lattices development corpus D and a 419 lattices test corpus T. Furthermore, the test corpus T has been split in two sets: T^+ which contains the 280 lattices for which the correct transcription is a sequence of words that corresponds to an existing path in the lattice and T^- which contains the 120 remaining lattices. We will present the results according to these two corpus because our aim is to precisely evaluate the gain of our rescoring method when all the information is available in the lattice produced by the first decoding process. Table 1 shows the baseline performances of B on the three corpora T, T^+ and T^-.

corpus	T	T^+	T^-
WER	27.12	11.26	50.81

Table 1. Results of the baseline system B

The poor results obtained on T^- lead us to consider that, with more than 50% WER, these sentences should be rejected by the dialogue system. For this reason we will focus

on the results obtained on T^+.

4. DATA AUGMENTATION BASED ON DISTANCES BETWEEN WORD REPRESENTATION

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_j. Let a_{ij} be the count for the same word and history, but after data augmentation. Let $\Gamma_j^\alpha(\theta)$ be the set of vectors representing the histories having distance lower than a threshold θ from the vectors representing h_j in the reduced space S_α. Let d_{jk}^α be the distance between vectors representing histories h_j and h_k in reduced space S_α. The augmented count a_{ij} of the sequence $h_j w_i$ is obtained assuming a history h_k *similar* to h_j contributes to the counts of the sequence $h_j w_i$ in a way that depends on a degree of similarity between the two histories h_j and h_k:

$$a_{ij} = c_{ij} + \sum_{h_k} c_{ik} \times f(d_{ik}^\alpha) \tag{4}$$

with: $h_k \in \Gamma_j^\alpha(\theta)$

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

$$f(d_{ik}^\alpha) = e^{-\frac{d_{ik}^\alpha}{D}} \tag{5}$$

where D is a parameter that can be used for tuning the system. The Euclidian distance between each pair of history vectors was computed in reduced space. The angle between each pair of history vector was also considered because it is used in Information Retrieval but it was abandoned because it produced poor results.

A first experiment was performed using a function defined by equation 5 with $D = 1$ with contribution from all histories. Results did not show a strong improvement with data augmentation, but suggested that history dependent thresholds should be used. A number of different ideas was considered. It was found that augmenting each bigram count with contributions only from the *K nearest histories* was effective. It may happen, that, for a given K, there are many histories with very close distances w.r.t. the K-th one. Consequently, 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*.

The parameters K and D have been calculated on the development corpus, and the following values have been chosen: $K = 77$, $D = 1$. The results are presented in table 2.

corpus	T	T^+	T^-
WER	26.95	10.53	51.14
gain	+0.6%	+6.5%	−0.6%

Table 2. Results of the quasi-K nearest histories augmentation: A_1

With these types of count adaptation, 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 for the sake of brevity.

5. ADAPTATION WITH WEIGHED COUNT AUGMENTATION ON SELECTED HISTORIES

In [1], 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 $c_g(w_i, w_j)$ $c_d(w_i, w_j)$ be the bigram counts, respectively, in the general corpus and in the domain adaptation corpus. An adaptation can be performed by interpolation of the general model counts and the counts of the adaptation corpus.

Table 3 shows the results obtained by applying this method to all the histories of an LM trained with the corpus from *Le Monde* (g) using the training corpus of domain specific (d) data for adaptation.

corpus	T	T^+	T^-
WER	26.95	11.04	50.38
gain	+0.6%	+1.9%	+0.8%

Table 3. Results of the global interpolation: A_0

Even if the results are better, the gain observed is rather small. Nevertheless, larger improvements have been observed by using counts from some specific histories of g to augment counts for the same histories of d. In fact, adapted and augmented counts can be expressed as follows:

$$c_a(w_i, w_j) = \alpha(w_j) \times c_g(w_i, w_j) + c_d(w_i, w_j) \tag{6}$$

The same type of data augmentation is applied to the counts of all words having a given history, while the type

of augmentation proposed in the previous section was applied to each word count, using contributions from different histories. The two types of augmentation are thus complementary.

Selecting the histories to be used for augmentation is crucial. There are histories represented, for example, by pronouns, for which the words which may follow have to satisfy certain constraints, like, in French gender and number. It is possible that some of the words satisfying the constraints are not observed in the d corpus, but are present in the g corpus. These words should have a probability higher than the one obtained with back-off methods, but somehow lower than what they have in g, because the words may not be semantically as relevant in d as they are in g. Histories having a large chance of inducing constraints on the following words have to be selected and the weight $\alpha(w_j)$ of count augmentation induced by these histories must be suitably chosen and has to be history dependent.

To choose these histories, we calculated a new count matrix for each history j. In this matrix, we combine the counts of the history j from the corpora d and g by means of equation 6. The parameter $\alpha(w_j)$ was set constant and equal to the ratio of words in the domain and the general corpus.

Then, an LM was calculated on each matrix and a decoding process was performed. Every history j which leads to a gain in WER on the development corpus was stored in a list L_h. At the end of this process, we performed a new adaptation between the counts of M_d and M_g only for the histories stored in L_h. In this process, a value $\alpha(w_j)$ was calculated for each history j on the development corpus. By combining this adaptation with the one presented in section 4, we obtained the results A_2 which are shown in table 5.

The results are significantly better on T^+, even if a degradation is observed on T and T^+. We believe that improving reliable hypotheses is more important that degrading results on sentences which will be rejected anyway by the dialogue module.

corpus	T	T^+	T^-
WER	27.56	9.94	53.55
gain	-1.6%	$+11.7\%$	-5.3%

Table 4. Results of the augmentation method A_2

6. CONCLUSION

A new method for data augmentation has been proposed. It is based on distances between word representations in a reduced space. The method leads to better LMs with which lower WER are obtained. The improvement in the LM may not always lead to a big reduction in WER because its impact depends on the quality of the acoustic scores. Nevertheless, data augmentation and count reuse makes it possible to train different LMs for different dialogue situations with greater accuracy. These LMs can be used in a rescoring phase and new decision algorithms can be conceived based on the results obtained with different LMs. Decision making in presence of hypotheses generated by different, competing models applied just on a trellis of hypotheses emerges now as a promising research direction.

7. REFERENCES

[1] Federico M. and De Mori R., *Language Model Adaptation*, K. Ponting ed. Springer-Verlag, Berlin, New York, 1999.

[2] Seymore K. and Rosenfeld R., "Using story topics for language model adaptation," in *Proc. Eurospeech 97, Rhodes,Greece*, 1997.

[3] Besling S. and Meier H.G., "Language model speaker adaptation," in *Proc. Euospeeech95 pp. 1755-1758, Madrid, Spain*, 1995.

[4] Iyer R. and Ostendorf M., "Modeling long distance dependence in language: topic mixtures vs. dynamic cache models," in *IEEE Transactions on Speech and Audio processing SAP-7(1):30-39*, 1999.

[5] Chen S.F., Seymore K., and Rosenfeld R., "Topic adaptation for language modeling using unnormalized exponential models," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, Seattle WA*, 1998.

[6] Janiszek D., de Mori R., Bechet F., Matrouf D., and Mokbel C., "New language model adaptation algorithm based on the definition of cardinal distance," in *ESCA Workshop on Interactive Dialogue in Multi-Modal Systems, Kloster Irsee, Germany*, 1999.

[7] Bellagarda J, "Multi-span statistical language modeling for large vocabulary speech recognition," *IEEE Transactions on Speech and Audio processing SAP-6(5):456-467*, 1998.

[8] Berry M.W., "Large-scale sparse singular value computations," in *Int. J. Supercomp. Appl. Vol6, No 1, pp13-49*, 1992.

[9] Sadek D., Ferrieux A., Cozannet A., Bretier P., Panaget F., and Simoni J., "Effective human-computer cooperative spoken dialogue: the ags demonstrator," in *ICSLP'96, USA*, 1996.