QWI: A METHOD FOR IMPROVED SMOOTHING IN LANGUAGE MODELLING*

G. Bordel^1, I. Torres^1 and E. Vidal^2

^{1}Dpto. Electricidad y Electrónica. Universidad del País Vasco.
Apdo. 644 - 48080 Bilbao. SPAIN.
^{2}Dpto. Sistemas Informáticos y Computación. Universidad Politécnica de Valencia.
Apdo. 22012 - 46071 Valencia. SPAIN.

ABSTRACT

N-grams have been extensively and successfully used for Language Modelling in Continuous Speech Recognition tasks. On the other hand, it has been recently shown that k-testable Stochastic Languages (k-TS) are strictly equivalent to N-grams.

A major problem to be solved when using a Language Model is the estimation of the probabilities of events not represented in the training corpus, i.e. unseen events. The aim of this work is to improve other well established smoothing procedures by interpolating models with different levels of complexity (Quality Weighted Interpolation - QWI).

The effect of QWI was experimentally evaluated over a set of back-off smoothed k-TS language models. These experiments were carried out over several corpora using the test-set perplexity as an evaluation criterion. In all the cases the introduction of QWI resulted in a reduction of the test-set perplexity.

1. INTRODUCTION

It is broadly accepted that large-vocabulary and/or Continuous Speech Recognition (CSR) Systems require a Language Model to integrate the syntactic and/or semantic constraints of the language. Language Modelling (LM) methodologies are often classified into two main categories: *Syntactic* and *Statistical*. Language constraints could be better modelled under a Syntactic approach, i.e. regular and/or context free grammars. However, they present computational complexity problems leading to the use of only very simple and restrictive models. As a consequence statistical methods are often preferred. They are based on the estimation of the probability of observing a given lexical unit, conditioned on the observations of N preceding lexical units (N-gram models). The number of probabilities to be taken into account is, in principle, an exponential function of N. In such a case, this formulation is only able to represent very local constraints and, as a consequence, does not model in an adequate way the inherent redundancy of the language [1]. Alternatively, high values of N have been proposed elsewhere [2].

On the other hand, it has been recently shown that k-testable Stochastic Languages (k-TS) are strictly equivalent to N-grams [3]. At this point, the above mentioned syntactic/stochastic classification may be questioned and then, choosing k-TS or N-grams could be just a matter of representation convenience [4].

A major problem to be solved when using a Language Model is the estimation of the probabilities of events not represented in the training corpus, i.e. *unseen events*. This problem has been extensively discussed resulting in some well known smoothing methods: linear and non linear interpolation [5] [6], coocurrence smoothing [7], back-off smoothing [8], etc.

This work provides an alternative method to deal with these problems called *Quality Weighted Interpolation* (QWI) (Section 3). The aim of the method is to improve other well established smoothing procedures by interpolating models with different levels of complexity. Therefore the method is applied to families of language models in which the same events (e.g. transitions in the case of Finite Automata) may be observed at the different complexity levels (*k*) of the family. We then chose the *K*-TS languages framework to apply the QWI smoothing method. Thus we first summarizes in Section 2 the grammatical formalism for N-gram models previously introduced in [9]. The classical *back-off* smoothing procedure [8] is also reformulated under this approach [9].

The effect of QWI was experimentally evaluated over a set of back-off smoothed k-TS language models. These experiments were carried out over several corpora using the test-set perplexity as evaluation criterion. The proposed formalism was compared with both the classical [8] and the previously proposed syntactical [9] back-off smoothing (Section 4). In all the cases the introduction of QWI resulted in a reduction of the test-set perplexity

2. N-GRAM LANGUAGE MODELLING UNDER A GRAMMATICAL FORMALISM

In an N-gram Language model, the probability of observing a given lexical unit, ω_j conditioned on the observations of *n* preceding lexical units, $P(\omega_i|\omega_{i-k+1}...\omega_{i-1})$, is estimated by counting the number of occurrences of each string $\omega_{i-k+1}...\omega_i$ in a given training corpus. The probability assigned to a new sentence or string of words is the product of the probabilities of the N-grams that appear in this sentence.

In spite of the many efforts to find alternatives to this approach, the most successful language models have actually been implemented according to this technique [10]. Nevertheless, the models obtained by the N-gram technique constitute in fact a proper (small) subset of the set of Stochastic Regular Grammars. It has been recently demonstrated that N-grams are strictly equivalent to "K-Testable in the Strict Sense" (K-TSS) Languages [11] [3]. Moreover, an inference algorithm to obtain Stochastic Finite Automata accepting K-TSS Languages was also developed [11]. These automata are deterministic and hence unambiguous [11]. Examples of K-TSS automata are shown in Figure 1.

* This work has been partially supported by the Spanish CICYT under grant TIC94-0210-E

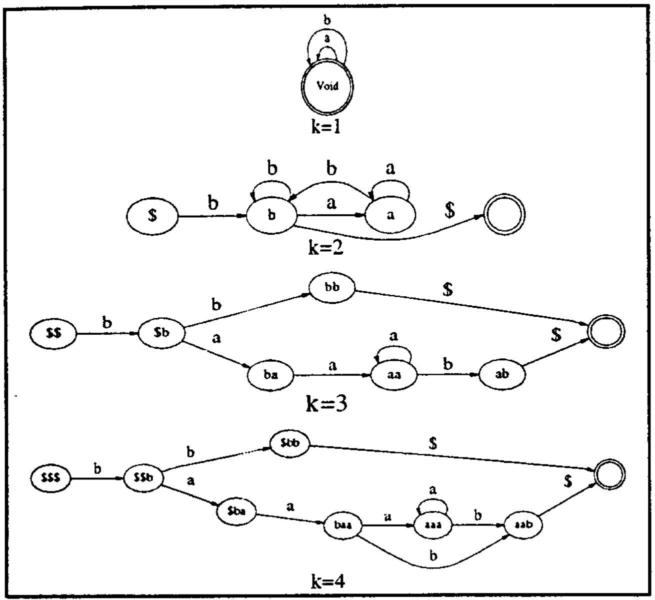

Figure 1: K-testable automata learned from the sample {'bb', 'baab', 'baaab', 'baaaab'}.

Given the unambiguous nature of any *K*-gram automaton the probability assigned to a sentence $\Omega \equiv \omega_1 ... \omega_l$ of length *l* is obtained as the product of the transitions used to accept Ω:

$$P(\Omega) = \prod_{i=1}^{l} P(\omega_i | \omega_{i-K+1}^{i-1})$$

where $P(\omega_i | \omega_{i-K+1}^{i-1})$ is the probability of a transition $\delta_K(E, \omega_i, E')$ from the state $E = \omega_{i-K+1}^{i-1}$ to $E' = \omega_{i-K+2}^{i}$ with the word ω_i.

The unambiguity of the Automaton also allows to obtain a simple maximum likelihood estimation of the probability of each transition [12]:

$$\hat{P}(\omega_i | \omega_{i-K+1}^{i-1}) = \frac{c(\omega_i | \omega_{i-K+1}^{i-1})}{\sum_{\omega_j \in \Sigma} c(\omega_j | \omega_{i-K+1}^{i-1})}$$

where $c(\omega_j | \omega_{i-K+1}^{i-1})$ is the number of times the word ω_i appears at the end of the *K*-gram $\omega_{i-K+1} ... \omega_{i-1}\omega_j$; i.e. the count of the transition labelled ω_i coming from state ω_{i-K+1}^{i-1}.

In classical smoothing techniques a specific probability mass is reserved to be shared among unseen events, i.e. *K*-grams not appearing in the training corpus. In Back-off smoothing [8] the probability to be assigned to unseen *K*-grams is recursively obtained from less accurate models, i.e. *K-1, ...,1*. At each step of the recursion the Turing formula [13] is used to estimate the discounted probability. Due to the drawbacks of the Turing formula in a practical application, the Back-off technique has been later tested under alternative estimations for the discount [6] [14]. In any case the same discount is established for all *K*-grams having the same counts.

The syntactic formalism suggests a state-dependent estimation of the total discount. In this case a counter is incremented by one each time a new word appears as an outgoing transition. This value represents the probability to be assigned to the "set of unseen words". The first time a new word is seen in a given state the new transition is established and its counter initialized to one [9].

In consequence, a set of transitions corresponding to the seen *K*-grams and one transition representing the set of unseen *K*-grams can be considered at each state. Let $E \equiv \omega_{i-K+1}^{i-1} \in Q_K$ be an state of the finite automaton $(\Sigma, Q_K, q_{0K}, q_{fK}, \delta_K)$ representing a *K*-gram model and $e \equiv \omega_{i-K+2}^{i-1} \in Q_{K-1}$ the corresponding state of the *(K-1)*-gram model. Let $u_E \equiv \Sigma - \Sigma_E$, where $\Sigma_E \equiv \{\omega_j | c(\omega_j | E) \neq 0\}$, represent the set of unseen *K*-grams at state *E*. Considering a maximum likelihood estimation the probability redistribution can be formulated as:

$$P(\omega_j | E) = \begin{cases} \left.\begin{cases} \dfrac{c(\omega_j | E)}{C_E} & \text{if } c(\omega_j | E) \neq 0 \; : \\ \dfrac{c(u_E | E)}{C_E} \dfrac{P(\omega_j | e)}{1 - \sum_{\forall \omega_r \in \Sigma_E} c(\omega_r | e)} & \text{if } c(\omega_j | E) = 0 \end{cases}\right\} & \text{if } E \in Q_k \\ P(\omega_j | e) & \text{if } E \notin Q_k \end{cases}$$

where $C_E = c(u_E | E) + \sum_{\forall \omega_r \in \Sigma_E} c(\omega_r | E)$.

3. QWI SMOOTHING

The aim of the method is to improve other well established smoothing procedures by interpolating models with different levels of complexity. The contribution of each individual model is calculated on the basis of its "quality". Thus, different language models need to be previously established, possibly using classical methods like flat smoothing or back-off smoothing.

According to QWI, given two models with complexity *K* and *K-1*, the probability of observing a given event ζ is calculated as:

$$P^*_k(\zeta) = \lambda_k P_k(\zeta) + (1 - \lambda_k) P^*_{k-1}(\zeta^*) \qquad (1)$$

where $P_k(\zeta)$ is the probability assigned to the same event by the corresponding classical model, $P^*_{k-1}(\zeta^*)$ represents the probability obtained by QWI for the same event in the *K-1* model and λ_k is the interpolation coefficient.

This formulation can be applied to families of language models in which the same events, i.e. transitions in the case of Finite Automata, may be observed at the different complexity levels (*K*) of the family. Within the N-Gram or k-TS languages framework:

$$P^*_k(\omega_j | E) = \lambda_k P_k(\omega_j | E) + (1 - \lambda_k) P^*_{k-1}(\omega_j | e) \qquad (2)$$

These formulae constitute a linear interpolation smoothing procedure as proposed by other authors [15]. Here, however, $P^*_k(\omega_i | E)$ is recursively obtained from less accurate models as in the back-off smoothing [8].

On the other hand, in QWI the calculation of the coefficients λ_k *k=1, 2, ...* is performed in such a way that the contribution of each model is weighted by its quality. The previously obtained models need thus to be evaluated over a specific validation set *T*. Using the inverse of the test-set perplexity (*PP*) as evaluation criterion, λ_k can be computed as:

$$\lambda_k = \frac{PP^*_{k-1}(T)}{PP_k(T) + PP^*_{k-1}(T)} \qquad (3)$$

By using (3) in (2), a set of recursive equations is obtained as:

$$P'_k(\omega_j|E)=\frac{PP'_{k-1}(T)P_k(\omega_j|E)+PP_k(T)P'_{k-1}(\omega_j|e)}{PP_k(T)+PP'_{k-1}(T)} \quad (4)$$

In the base of the recursion (K = 1), the model consists of only the void state (see Figure 1). Thus $P_1(\omega_j|E)$ becomes $P_1(\omega_j)$ and:

$$P'_1(\omega_j|E)=P'_1(\omega_j)=\frac{PP'_0(T)P_1(\omega_j)+PP_1(T)P'_0}{PP_1(T)+PP'_0} \quad (5)$$

where P'_0 and $PP'_0(T)$ represent the probability and the perplexity corresponding to a null model (K=0) and are calculated as:

$$P'_0=P_0=\frac{1}{|\Sigma|} \qquad PP'_0(T)=PP_0(T)=|\Sigma| \quad (6)$$

where $|\Sigma|$ is the size of the vocabulary.

If a given fixed training set R is available, we can partition R into training and validation sets in many ways. Obviously, better estimates can be obtained by performing a cross-validation type partitioning procedure [16]. On the other hand, once the QWI values of λ_k have been obtained, we can use the corresponding k-TS models to evaluate the validation-set perplexities. In this way, the whole process can be iterated until stable, and hopefully "optimal", values for λ_k are achieved. Iterating Equation (4) until convergence:

$$P''_k(\omega_j|E)=\frac{PP'_{k-1}(T)P_k(\omega_j|E)+PP'_k(T)P'_{k-1}(\omega_j|e)}{PP'_k(T)+PP'_{k-1}(T)} \quad (7)$$

The base of the recursion is now:

$$P''_1(\omega_j|E)=\frac{|\Sigma|P_1(\omega_j|E)+PP'_1(T)\frac{1}{|\Sigma|}}{PP_1(T)+|\Sigma|} \quad (8)$$

where PP' is the perplexity value obtained through the previously smoothed model.

4. EXPERIMENTAL RESULTS

The effect of QWI was experimentally evaluated over a set of back-off smoothed k-TS language models. Both back-off proposed by Katz [8] and back-off recently introduced by the authors in the syntactic framework [9] were considered.

These experiments were carried out over three different corpora consisting in 9150 sentences each one. However the task difficulty is not the same. Moreover, the sentence length, total number of words and size of the vocabulary are different in each case. As a consequence, different perplexity values were obtained. The first one consists of a set of simple English sentences describing visual scenes (Miniature Language Acquisition task – MLA) [17]. The required number of sentences were randomly generated by using a context-free model of the language [17]. It includes 147002 words and a very limited vocabulary size (29 words). Thus in this case the back off proposed by Katz could not be applied since the Turing formula assumes a certain probability distribution which is only guaranteed with wide vocabularies. The second corpus (BDGEO) is a task-oriented Spanish speech corpus [18] consisting in 82000 words and a vocabulary of 1284 words. This corpus consists of a set of Natural Language (spontaneous) queries to a Spanish geographic database. This is a very specific task designed to test integrated systems (acoustic, syntactic and semantic modelling) in automatic speech understanding, which leads to low perplexity values. The third corpus is a subset of the English transcription of the Bible (BIB). This is the most difficult corpus including 255380 words and having a vocabulary size of 8122 words.

A cross-validation technique was applied [16] selecting in each random partition 9100 sentences for training purposes and 50 sentences for testing. In each case a validation subset of 300 sentences was selected from the training set to obtain the λ_k coefficients required by QWI. In all the cases the test-set perplexity was used as evaluation criterion.

Figures 2, 3 and 4 show the test set perplexities for the experiments carried out over the MLA, the BDGEO and the BIB corpus respectively. In the first case (Figure 2, MLA task) only the syntactic approach was considered. The minimum perplexity value was obtained at a high value of k (7) the relatively large amount of training data with respect to the complexity of the language. In this case the introduction of QWI did not improve this value. In Figure 3 (BDGEO) both back-off approaches were improved by the introduction of QWI. The best perplexity value was obtained by the syntactic back-off with QWI and k = 6. Finally, both classical and syntactical back-off were improved by the introduction of QWI when the BIB corpus was used (Figure 4). Due to the high computational cost of this experiment, Figure 4 only represents one partition (not the entire cross-validation process). So, it should be interpreted in comparative terms (the values for the perplexity may slightly change at the end of the process). The relationship between the training size and the size of the vocabulary lead to obtain the minimum perplexity at a lower value of k (5).

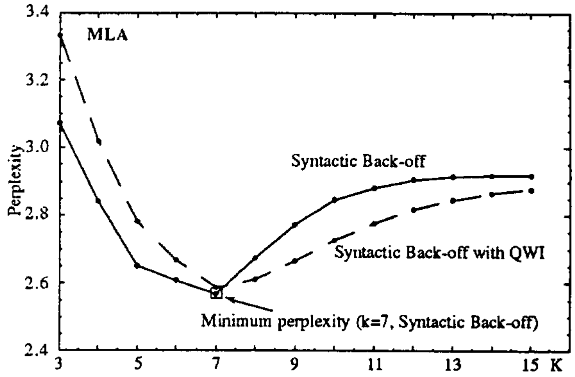

Figure 2: Test set perplexities for k = 3, ..., 15 obtained trough the experiments carried out over the MLA corpus.

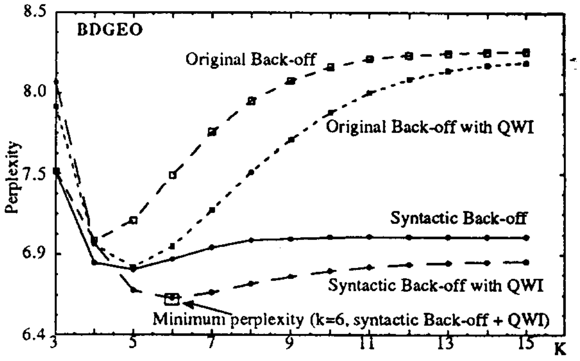

Figure 3: Test set perplexities for k = 3, ..., 15 obtained trough the experiments carried out over the BDGEO corpus.

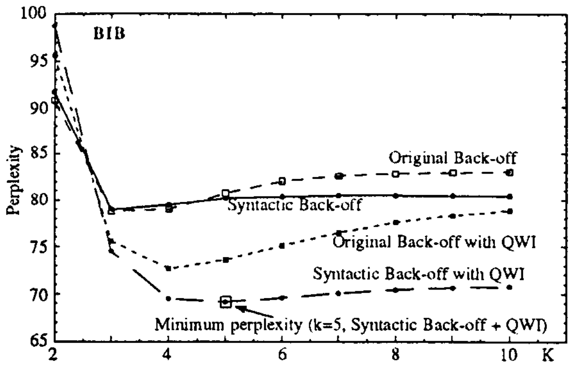

Figure 4: Test set perplexities for k = 3, ..., 10 obtained trough the experiments carried out over the BIB corpus

5. CONCLUSIONS

The aim of this work was to improve other well established smoothing procedures by interpolating models with different levels of complexity. Therefore the method was applied within the K-TS languages framework which are strictly equivalent to N-grams. Both the back-off proposed by Katz and the back-off recently introduced by the authors in the syntactic framework were chosen to be improved by QWI.

The experimental evaluation was carried out over three corpora including different sentence length, total number of words, vocabulary size and task difficulty. The introduction of QWI resulted in a reduction of the test-set perplexity for values of k higher than a certain threshold. This reduction, usualy lead to obtain the minimum perplexity value. The value of k at which this minimum was obtained was strongly related with the relationship between the training size and the size of the vocabulary. Thus for the largest vocabulary size (BIB) this value is lower than for the smallest one (MLA) when the same number of training sentences was considered. On the other hand the significance of the improvement introduced by QWI was more important for large vocabulary sizes.

Regardless of the smoothing method there is always a value of K from which the perplexity value began to increase when K did. However this effect was less important and appeared at higher values of K for QWI, specially within the syntactic framework.

Finally let us to note that even the number of probabilities to be calculated in an N-gram model is, theoretically, an exponential function of N, this function is in practice attenuated very early. Moreover, the global automaton required by the syntactic approach leads to a very fast procedure to get sentence probabilities and provides a good expectation of integration with the acoustic models in real speech understanding tasks.

REFERENCES

[1] H. Ney, "Architecture and Search Strategies for Large-Vocabulary Continuous-Speech Recognition," *NATO ASI. New Advances and Trends in Speech Recognition and Coding*, Lectures, pp. 59-84. Bubión, 1993.

[2] G. Bordel, "Modelización del lenguaje: una visión general desde el análisis de los lenguajes K-explorables en sentido estricto (N-gramas)," Research report DSIC-II/40/93 DEE-1/93, Universidad Politécnica de Valencia, Universidad del País Vasco, 1993.

[3] E. Segarra, "Una Aproximación Inductiva a la Comprensión del Discurso Continuo," Ph Thesis. Universidad Politécnica de Valencia, 1993.

[4] E. Vidal, F. Casacuberta and P. García, "Syntactic Learning Techniques for Language Modelling and Acoustic-Phonetic Decoding," *Lectures of the NATO ASI* (New Advances and Trends in Speech Recognition and Coding), pp. 95-201, Bubión jun-jul 1993.

[5] F. Jelinek, "Markov Source Modelling of Text Generation," *The impact of processing techniques on communication.*, Ed. J.K.Skwirzynski, Nijhoff, Dordrecht, The Netherlands, 1985.

[6] H. Ney, U. Essen, "On Smoothing Techniques for Bigram-Based Natural Language Modelling," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 825-828, 1991.

[7] U. Essen and V. Steinbiss, "Coocurrence Smoothing for Stochastic Language Modelling," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 161-164, 1992.

[8] S. M. Katz, "Estimation of Probabilities from Sparse Data for The Language Model Component of a Speech Recognizer," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-35, nº.3, pp.400-401, march 1987.

[9] G. Bordel, I. Torres and E. Vidal. "Back-off Smoothing in a Syntactic approach to Language Modelling". *International Conference of Spoken Language Processing* Japan, September 1994.

[10] F. Jelinek, "Up from trigrams: the struggle for improved language models," *proc. of the Eurospeech 91*; Genova, Italy, pp.1037-1039, sep. 24-26, 1991.

[11] P. García and E. Vidal, "Inference of k-testable languages in the strict sense and application to syntactic pattern recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol 12, nº 9, pp. 920-925, sep. 1990.

[12] R. Chandhuri and T. L. Booth, "Approximating Grammar Probabilities: Solution of a Conjecture," *Journal ACM*, vol 33, nº 4, pp. 702-705, 1986.

[13] I. J. Good, "The Population Frequencies of Species and the estimation of population parameters," *Biometrika*, vol. 40, parts 3 and 4, pp. 237-264, 1953.

[14] H. Ney, U. Essen and R. Kneser, "On Structuring Probabilistic Dependencies in Stochastic Language Modelling," *Computer Speech and Language* 8, pp 1-38, 1994.

[15] F. Jelinek. "Self-organized Language Modeling for Speech Recognition". IBM Europe Institute; Advances in Speech Processing. Oberlech, Austria. July, 1986.

[16] S. J. Raudys and A. K. Jain, "Small Sample Effects in Statistical Pattern Recognition: Recommendations for Practitioners and Open Problems," *IEEE Trans. on PAMI*, vol. 13, n3, pp. 252-263, 1991.

[17] J.A. Feldman, G. Lakoff, A. Stolcke, S.H. Weber, "Miniature Language Acquisition: a touch-stone for cognitive science". Technical report, TR-90-009. ICSI, Berkeley, California. April, 1990.

[18] J. E. Diaz, A. J. Rubio, A. M. Peinado, E. Segarra, N. Prieto and F. Casacuberta, "Development of Task Oriented Spanish Speech Corpora," *Proceedings of EUROSPEECH 93* , 1993.