Back-off smoothing evaluation over syntactic language models

A. Varona, I. Torres

Dpto. Electricidad y Electrónica. Universidad del País Vasco
Apdo. 644 48080 Bilbao. SPAIN

E-mail (amparo@we.lc.ehu.es, manes@we.lc.ehu.es)

Abstract[1]

Continuous Speech Recognition systems require a Language Model (LM) to represent the syntactic constraints of the language. In LMs development a smoothing technique needs to be applied to also consider events not represented in the training corpus. In this work, several back-off smoothing approaches have been compared: classical discounting-distribution schema including Witten-Bell, Absolute and Linear discounting and a new proposal, the Delimited discounting. Delimited discounting deals with the Turing discounting problems while keeping the Katz´s smoothing scheme. The experimental evaluation was carried out over a Spanish speech application task, showing that an increase of the test set perplexity of a LM does not always mean a degradation in the model performance when integrated into a CSR system. Besides, there is a strong dependence between the amount of probability reserved by the smoothing technique to be assigned to *unseen* events and the value of the balance parameter applied to the LM probabilities in the Bayes´s rule needed to get the best system performance.

1. Introduction

Continuous Speech Recognition (CSR) Systems require a Language Model (LM) to integrate the syntactic and/or semantic constraints of the language. In this work, a syntactic approach of the well-known n-gram models is used to get suitable LMs: the k-Testable in the Strict Sense (k-TSS) LM [1] which are a sub-class of regular machines. Conventional n-grams do not include any structural parameter whereas k-TSS automata is fully represented and stored [2]. Thus, choosing k-TSS or n-grams is just a matter of representation convenience [3].

LM are usually get from large text databases. Then a smoothing technique needs to be applied to also estimate the probabilities to be assigned to those events not represented in the training corpus, that is, *unseen* events [4] [5]. Backing-off smoothing was chosen in previous works [6] because the involved recursive scheme has been well integrated into the finite state formalism. In this work, several back-off smoothing approaches have been evaluated and compared: classical discounting-distribution scheme including Witten-Bell, Absolute and Linear discounting a new proposal, the Delimited discounting [7]. In Witten-Bell, Absolute and Linear discounting, discounting factors are applied to the whole set of seen events in the training corpus. As a consequence, the mass of probability to be assigned to unseen events could be overestimated. On the contrary, in the new Delimited proposal, discounting factors are only applied to those events scarcely observed in the training corpus. That is, it is based on the well-known Turing discounting [8] but avoiding its associate problems.

The evaluation was carried out using the test set perplexity and the obtained %WER in the CSR system. Nevertheless, the ability of the test set perplexity to predict the real behavior of a smoothing technique when working in a CSR system could be questioned [9] because it has not take into account the relationship with acoustic models.

Usually, heuristic parameters are applied to the Bayes´ rule to obtain optimum CSR performances [10] [11]. In this work, the effect of applying a balance exponential factor to the LM probabilities was also evaluated. This evaluation shows a strong relationship between the amount of probability reserved by the smoothing technique to be assigned to *unseen* events and the value of the scaling factor required to obtain the best CSR system performance.

Section 2 deals with back-off smoothing proposals to be evaluated: the classical Witten-Bell, Absolute Linear and the new Delimited discounting. In section 3 the experimental evaluation of the smoothing techniques was evaluated in terms of both, perplexity and *WER*. Finally, some concluding remarks are showed in Section 4.

2. Back-off smoothing techniques

As mentioned above, the *k*-Testable in the Strict Sense (*k*-TSS) languages are a sub-class of the regular languages. The use of regular grammars allowed to obtain a deterministic *k*-TSS stochastic finite state automaton that can be efficiently integrated in a CSR system [1] [3]. In such a model, each state of the automaton represents a string of words $w_{i-(k-1)}w_{i-(k-1)}...w_{i-1}$ and it is labeled as $w_{i-(k-1)}^{i-1}$, where *i* stands for a generic index in any string $w_1...w_i...$ appearing in the training corpus. Each transition represents a *k*-gram, it is labeled by its last word w_i and connects two states labeled up to with *k*-1 words.

The probability associated with each transition representing *seen events* can be estimated under a maximum likelihood criterion. However, a probability needs also to be associated with *unseen events*. To deal with this problem, backing-off smoothing was chosen in previous works [6] because the involved recursive scheme has been well integrated in the finite state formalism. The modified probability $P(w/q)$ to be associated with a

[1] Work partially supported by the Spanish CICYT under grant TIC98-0423-C06-03

transition is estimated according to:

$$P(w/q)=\begin{cases}[1-\lambda]\dfrac{N(w/q)}{N(q)} & w\in\Sigma_q\\ \left(\sum\limits_{\forall w_i\in\Sigma_q}\lambda\dfrac{N(w_i/q)}{N(q)}\right)\dfrac{P(w/b_q)}{1-\sum\limits_{\forall w_i\in\Sigma_q}P(w_i/b_q)} & w\in\Sigma-\Sigma_q\end{cases}\tag{1}$$

where: Σ is the vocabulary of the task, that is, the set of words appearing in the training corpus; Σ_q is the vocabulary associated with state *q* and consists of the set of words appearing after the string labeling state *q*, i.e. words labeling the set of *seen* outgoing transitions from state *q*; *N*(w/*q*) is the number of times that word w_i appears after the string labeling state *q*; $N(q)=\sum_{\forall w\in\Sigma_q}P(w/q)$,

and $P(w/b_q)$ is the estimated probability associated with the same event in the (*k*-1)-TSS model. In this schema, (1-λ) represents the discount factor, that is, the amount of probability to be subtracted and then be redistributed among *unseen* events. Figure 1 represents this schema for a state q labeled as $w_{i-(k-1)}^{i-1}$.

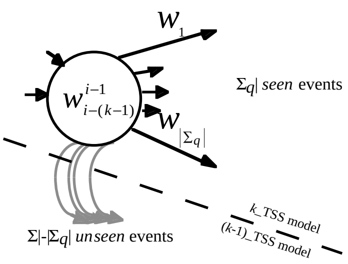

Figure 1: |Σ_q| seen events and |Σ|-|Σ_q| unseen events can be found at each state *q* labeled as $w_{i-(k-1)}^{i-1}$. The probability associated with *unseen* events is recursively obtained from less accurate models (*k*-1, *k*-2,…1) using Equation 1.

2.1.- Discounting over all seen events

In the first place, the discounting factor (1-λ) can be applied to the whole set of seen events in the training corpus, as it is suggested in Equation 1. Absolute and Linear discounting are classical proposals in which the discounting factors depend on respective parameter values, whereas the Witten-Bell discounting does not depend on any parameter.

Witten-Bell discounting

The discounting depends on the number of different events following the particular context labeling a state *q*, i.e. the size of the state vocabulary |Σ_q|, and on the number of seen events appearing at this context *N*(*q*):

$$1-\lambda=\frac{N(q)}{N(q)+\left|\Sigma_q\right|}\tag{2}$$

Absolute discounting

This discounting schema consists on subtracting a constant *b* from each count *N*(*w*/*q*) in the following way:

$$1-\lambda=\frac{N(w/q)-b}{N(w/q)}\tag{3}$$

Linear discounting

In this case a quantity proportional to each count is subtracted from the count itself in the following way:

$$1-\lambda=1-l\tag{4}$$

2.2.- Discounting over scarcely seen events.

On the other hand, discounting factor could be only applied to the scarcely observed seen events: Katz discounting. The scheme devised by Katz [8] combines Turing discounting with backing-off. According to this formalism the probability associated with events occurring more than a fixed number of times, say *r* times, are estimated under a maximum likelihood criterion whereas events occurring less than *r* times, $N(w_i/q)<r$, are discounted a certain mass of probability. Thus:

$$P(w/q)=\begin{cases}\dfrac{N(w/q)}{N(q)} & w\in\Sigma_q\wedge N(w/q)>r\\ [1-\lambda]\dfrac{N(w/q)}{N(q)} & w\in\Sigma_q\wedge 1\le N(w/q)\le r\\ \sum\limits_{\substack{\forall w_i\in\Sigma_q\\ 1\le N(w_i/q)\le r}}\left[\lambda\dfrac{N(w_i/q)}{N(q)}\right]\dfrac{P(w/b_q)}{1-\sum\limits_{\forall w_i\in\Sigma_q}P(w_i/b_q)} & w\in\left(\Sigma-\Sigma_q\right)\end{cases}\tag{5}$$

In Turing discounting the discounted mass of probability depends on n_1, n_2, ...,n_{r+1}, (being n_i the number of events which occur *i* times). The lower the count *N*(*w*/*q*) is, the bigger discounting is applied, because higher counts are supposed to be better estimated. This approach puts some constrains to the relative values of n_1, n_2, ...,n_{r+1}, which are not always satisfied by *k*-grams models with medium and high values of *k*, due to the lack of an adequate distribution of the samples.

Delimited discounting

To avoid the Katz discounting problems, the Delimited discounting [7] can be used. As in the Katz model, the discounting operation was limited to low counts, i.e., $N(w/q)\le r$ in the following way:

$$1-\lambda=d-\tau(r-N(w/q))\qquad \tau,d<1\wedge\tau<<<d\tag{6}$$

Discounting depends on *d* and τ parameters' values, which must be minor than one. The bigger the count was ($N(w/q)\le r$) the lower discounting was applied. When *N*(*w*/*q*)=*r*, the discounting was the minimum (only depends on *d* parameter), and when *N*(*w*/*q*)=1, the discounting applied was the maximum (*d*-*t*(*r*-1)).

3.- Experimental evaluation.

Previously presented back-off smoothing techniques were evaluated and compared over a set of *k*-TSS language models in terms of test set perplexity and *WER* in a CSR system [12]. Since k-TSS LM are a syntactic representation of classical n-grams, the obtained results in this work are valid for both aproximations.

The experiments were carried out over a task-oriented Spanish speech corpus [13], consisting in 82,000 words and a vocabulary of 1,208 words. This corpus represents a set of queries to a Spanish geography database. The training corpus used to obtain the *k*-TSS models, consisted in 9150 sentences. The text test set consisted in 200 different sentences. These sentences were then uttered by 12 speakers resulting in a total of 600 sentences that composed the speech test set. Uttered sentences were decoded by the time-synchronous Viterbi algorithm with a fixed beam-search to reduce the computational cost. A chain of Hidden Markov models representing the acoustic model of the word phonetic chain replaced each transition of the *k*-TSS automaton.

In a first series of experiments test set perplexities (PP) were obtained. Absolute (Ad), Linear (Ld) and Delimited (Dd) discounting depend on values of *b*, *l*, and *d* parameters respectively. For each discounting method, two values were chosen to be evaluated: those which got a “good” LM (low PP) and a “bad” LM (high PP). Besides, in Delimited discounting τ parameter was fixed (τ=0.01) and the minimum number of times *r* required for a maximum likelihood estimation of event probabilities (Equation 5) was also set to $r\approx 7$. Witten-Bell (WBd) is not depending on any parameter.

Figure 2 shows the results of this first series of experiments. Techniques involving a low smoothing degree (Ad *b*=0.01, Dd *d*=0.99 and Ld *l*=0.01) achieved high perplexity values, whereas techniques involving a high smoothing degree (Ad b=0.4, WBd and Dd d=0.70) achieved low perplexity values, being constant for high values of *k*.

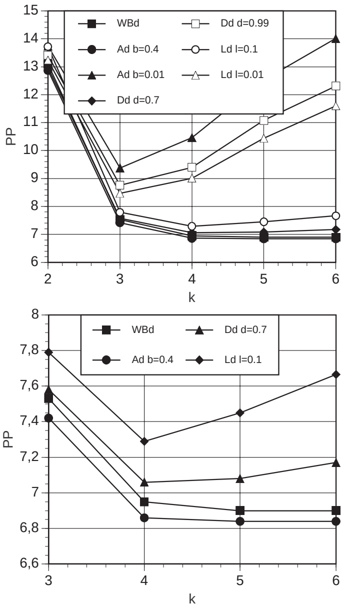

Figure 2. - a) PP obtained by several Smoothed *k*-TSS LM using Witten-Bell (WBd), Absolute (Ad), Linear (Ld) and Delimited (Dd) discounting methods. b) A detail of Figure 2a for PP values around 7.

Figure 3 shows the system performance (*WER*) obtained when these smoothed LMs were integrated in the CSR system. The average number of active nodes in the trellis (including acoustic and LM states) needed by every LM to decode a sentence is also represented in Figure 3.

Figure 3 shows that low-smoothing techniques (Ad *b*=0.01, Dd *d*=0.99 and Ld *l*=0.01) led to the best system performances: low *WER* and average number of active nodes. On the other hand, high-smoothing techniques led to the worst system performances: Witten-Bell and Linear discounting (*l*=0.1). Absolute discounting (*b*=0.4) and Delimited discounting (*d*=0.70) had almost the same behavior (it is very difficult to see the difference in Figure 3). In all discounting methods there was not difference among the results obtained by *k*-TSS models with values of *k* > 3.

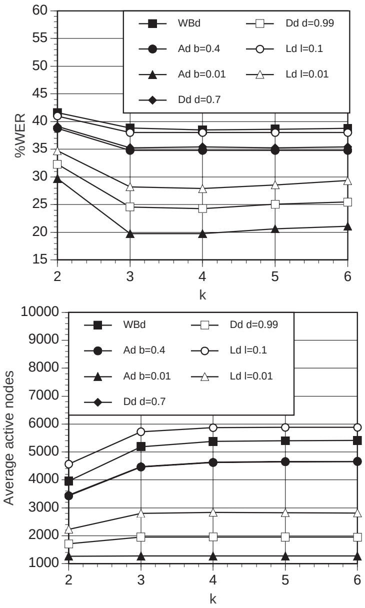

Figure 3. - a) System performance (*WER*) obtained for the Smoothed *k*-TSS LMs in Figure 2. b) Average number of active nodes at decoding time.

The word error rates shown in Figure 3 clearly disagree with the perplexity behavior shown in Figure 2. Moreover, the best system performances were obtained when smoothing methods leading the highest perplexity values were used [9].

In a second series of experiments the LM probabilities were modified by introducing a balance parameter α [10] [11] in the Bayer’ rule: $P(w)^{\alpha}$. Figure 4 shows the system performance (*WER*) obtained by two of the smoothed language models (*k*=3 and 4) used in the first series of experiments (Figures 2 and 3), when different values of the balance parameter around the optimum performance (α=3, 4, 5 and 6) were considered.

Points at the bottom left corner of each plot represent the best system performance: the lowest *WER* and the lowest average number of active nodes in the lattice. For any *k*-TSS model, important increases in both %WER rates (up to a minimum) and in the average active nodes can be observed (Figure 4) when the balance parameter α was increased.

The best performance for this task was reached by high-smoothing techniques: Witten-Bell with α=6 and Absolute (*b*=0.4) and Delimited (*d*=0.7) with α=5. Nevertheless, the *WER* differences around the optimum *WER* value were not significant. The system performance decreased when low-smoothing techniques were used in this task: Absolute (*b*=0.01), Delimited (*d*=0.99) and Linear (*l*=0.01) discounting. They reached their best system performance with α values around 3.

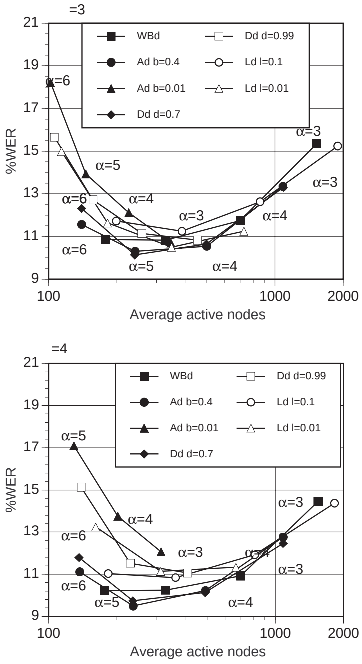

Figure 4. - %WER obtained by the Smoothed *k*-TSS LMs in Figure 2 and 3 using different values of the α parameter.

The accumulated probabilities at the end of each word Ω is a combination of acoustic P(A/Ω) and language P(Ω) probabilities. Acoustic probabilities are usually smaller than language probabilities and besides they are applied much more times. So that, the gap among accumulated probabilities is usually bigger than the gap among LM probabilities. The immediate consequence is that LM probabilities are irrelevant in most part of the situations to decide the best way to follow. However, when the LM probabilities are raised to a power $\alpha>1$: $(P(w))^{\alpha}$, all of them are attenuated, but this attenuation is higher for lower probability values. So that, the gap between the high and low probabilities is also bigger and then the LM probabilities are more and more competitive with the increase of α values, up to a maximum where LM probabilities are overvalued.

Since, high-smoothed LMs have a smaller gap among probabilities, they reached their best performance for values of α higher than low-smoothed LM. As a consequence, there is a strong dependence between the smoothing technique and the value of the scaling parameter α needed to get the best performance of the system (which is in many cases perplexity independent). In fact, the use of a balance factor α can be understood as a new smoothing (redistribution) of the LM probabilities.

4.-Concluding remarks.

Several back-off smoothing approaches have been tested and evaluated over syntactic Language Models. Classical discounting-distribution schema including Witten-Bell, Absolute and Linear discounting and a new proposal, the Delimited discounting, were evaluated and compared. Delimited discounting deals with the Turing discounting problems while keeping the Katz´s smoothing scheme.

The experimental evaluation was carried out over an Spanish application task using both, the test set perplexity and a CSR system performance (*WER*). Experiments show that an increase of the test set perplexity of a LM does not always means degradation in the model performance which fundamentally depends on empirical factors. In this work, several scalling factors were applied to the estimated smoothed LM probabilities. In all cases important decreases in both %WER and average active nodes in the lattice were observed. In fact, it was proved that there is a strong dependence between the amount of probability reserved by the smoothing technique to be assigned to *unseen* events and the value of the balance parameter α needed to get the best system performance.

5. References

[1] A. Varona and I. Torres (1999) "Using Smoothed K-TSS Language Models in Continuous Speech Recognition". *Procc. IEEE ICASSP*. pp.729-732.

[2] P. Clarkson, R. Rosenfeld. "Statistical language modeling using the CMU-CAMBRIDGE toolkit", (1997) *Proceedings of EUROSPEECH 97* pp- 2707-2710.

[3] I. Torres, A. Varona (2001): "k-TSS language models in a speech recognition systems". Computer Speech and Language. April Volume 15 No 2

[4] H. Ney, S., Martin, F Wessel, (1997): "Statistical Language Modeling using leaving-one-out". In S. Young and G. Bloothooft (eds.). Corpus-based methods in Language and Speech processing, pp. 174-207. Kluwer Academic Publishers

[5] S.F. Chen, J. Goodman, (1999). "An empirical study of smoothing techniques for language modeling". *Computer Speech and Language*. Vol 13. pp359-394

[6] G. Bordel, I. Torres and E. Vidal (1994): "Back-off smoothing in a syntactic approach to Language Modeling". *Proc.ICSLP-94*, pp. 851-854.

[7] A. Varona and I. Torres (2000): "Delimited smoothing technique over pruned and not pruned syntactic language models: perplexity and WER". Procc of the ISCA ITRW ASR 2000 Workshop, pp.69-76

[8] S. M.Katz. (1987). "Estimation of Probabilities from Sparce Data for The Language Model Component of a Speech Recognizer". *IEEE Trans. on Acoustics, Speech and Signal Processing*,. vol. ASSP-35, n 3, pp. 400-401.

[9] P. Clarkson, T. Robinson (1999). "Towards improved language model evaluation measures". *Procc of EUROSPEECH.99*. Vol 5. pp 1927-1930

[10] F. Jelinek, (1996): "Five speculations (and a divertimento) on the themes of H.Bourlard, H.Hermansky, N. Morgan". *Speech Communication* 18, pp 242-246

[11] A. Ogawa. K. Takeda, F. Itakura (1998) "Balancing Acoustic and Linguistic Probabilities". In Procc. ICASSP´98. Vol 1. pp 181-185.

[12] L.J. Rodriguez, I.Torres, J.M Alcaide. A. Varona, K. López de Ipiña, M.Peñagarikano. G. Bordel (1999). "An Integrated System for Spanish CSR Tasks". *Proc of EUROSPEECH.99*. Vol 2. pp 951-954

[13] J. Diaz, A. Rubio, M. Peinado, E. Segarra,, N. Prieto & F. Casacuberta (1993); "Development of Task Oriented Spanish Speech Corpora," Proc. EUROSPEECH 93.