Integrating High and Low Smoothed LMs in a CSR System*

Amparo Varona and Ines Torres

Departamento de Electricidad y Electrónica. Facultad de Ciencias. UPV/EHU.
Apartado 644. 48080 Bilbao. SPAIN.
{amparo, manes}@we.lc.ehu.es

Abstract. In Continuous Speech Recognition (CSR) systems, acoustic and Language Models (LM) must be integrated. To get optimum CSR performances, it is well-known that heuristic factors must be optimised. Due to its great effect on final CSR performances, the exponential scaling factor applied to LM probabilities is the most important. LM probabilities are obtained after applying a smoothing technique. The use of the scaling factor implies a redistribution of the smoothed LM probabilities, i.e., a new smoothing is obtained. In this work, the relationship between the amount of smoothing of LMs and the new smoothing achieved by the scaling factor is studied. High and low smoothed LMs, using well-known discounting techniques, were integrated into the CSR system. The experimental evaluation was carried out on two Spanish speech application tasks with very different levels of difficulty. The strong relationship observed between the two redistributions of the LM probabilities was independent of the task. When the adequate value of the scaling factor was applied, not very different optimum CSR performances were obtained in spite of the great differences between perplexity values.

1 Introduction

In Continuous Speech Recognition (CSR) systems a Language Model (LM) is required to represent the syntactic constraints of the language. But there are a high number of sequences of words that do not appear in training and could appear in tests. Thus, a certain mass of probability must be subtracted from the seen combinations and redistributed among the unseen ones, i.e., a smoothing technique must be applied [1] [2].

The test set perplexity is typically used to evaluate the quality of the LM [1] [2]. Perplexity can be interpreted as the (geometric) average branching factor of the language according to the model. It is a function of both the language and the model. It is supposed that the "best" models get the "lowest" Word Error Rates (WER) of the CSR system. But there are plenty of contraexamples in literature [3]. The ability of the test set perplexity to predict the real behavior of a

* This work has been partially supported by the Spanish CICYT under grant TIC2002-04103-C03-02 and by the Basque Country University under grant (9/UPV 00224.310-13566/2001)

A. Sanfeliu and J. Ruiz-Shulcloper (Eds.): CIARP 2003, LNCS 2905, pp. 236–243, 2003.

smoothing technique when working in a CSR system could be questioned because it does not take into account the relationship with acoustic models. Several attempts have been made to devise metrics that are better correlated with the application error rate than perplexity [4]. But for now perplexity remains the main metric for practical language model construction [3]. In fact, the quality of the model must ultimately be measured by its effect on the specific application for which it was designed, namely by its effect on the system error rate. However, error rates are typically non-linear and poorly understood functions of language models [3]. In this work we try to clarify how the smoothing technique applied to the LM works in the CSR system and to show its real impact on final system error rates.

Integration of language and acoustic models is invariably based on the well-known Bayes' rule. However, it is well known that the best performance of a CSR system is obtained when LM probabilities in the Bayes' rule are modified by introducing an exponential scaling factor [5] [6]. This factor can be understood as a new redistribution of the smoothed LM probabilities. As a consequence, LMs are smoothed twice: first by means of the smoothing technique and then by the exponential scaling parameter. The aim of this work is to establish a relationship between the amount of smoothing given by the smoothing technique and the amount of smoothing achieved by the exponential scaling factor (see Section 2).

Thus, different amounts of smoothing need to be applied to LMs. Two different well-stablished smoothing techniques leading to high and low-smoothed LM respectively, have therefore been evaluated (see Section 3). The relationship between the amount of smoothing given by the smoothing technique and the amount of smoothing achieved by the exponential scaling factor is studied in terms of both classical test set perplexity and CSR performance. CSR performance was evaluated in terms of both, the obtained WER and involved computational cost (see Section 4). Experimental evaluation was carried out over two Spanish databases of very different difficulty recorded by two consortia of Spanish research groups to work in understanding and dialogue systems respectively. Finally, some concluding remarks are given in Section 5.

2 Introducing the LM in the CSR System

Within a CSR system there are several heuristic parameters that must be adjusted to obtain optimum performances, such as the beam-search factor to reduce the computational cost, etc. But, the most important, due to its great effect on final CSR performance, may be the exponential scaling factor α applied over LM probabilities in Bayes' rule [5]. In Bayes' rule, the recognizer must find the word sequence $\hat{\Omega}$ that satisfies:

$$\hat{\Omega} = \arg\max_{\Omega} P(\Omega)^{\alpha} P(A/\Omega) \tag{1}$$

where $P(\Omega)$ is the probability that the word sequence $\Omega \equiv \omega_1\omega_2\ldots\omega_{|\Omega|}$ from some previously established finite vocabulary $\Sigma = \{\omega_j\}$, $j = 1\ldots|\Sigma|$, will be uttered and $P(A/\Omega)$ is the probability of the sequence of acoustic observations $A = a_1 a_2 ... a_{|A|}$ for a given sequence of words Ω. Probabilities $P(A/\Omega)$ are represented by acoustic models, usually Hidden Markov Models (HMM). The *a priori* probabilities $P(\Omega)$ are given by the LM.

From a theorical point of view, the scaling parameter α is needed because acoustic and LM probability distributions are not real but approximations [5]. The two probability distributions are estimated independently using different stochastic models that represent different knowledge sources. Moreover, the parameters of the acoustic and language models are estimated on the basis of speech and text data corpora, respectively. Each corpora was designed with different purposes, and they have therefore different vocabulary, size, complexity, etc. Thus, a balance parameter α needs to be applied to lessen these effects and then obtain good system performances.

In practice, acoustic and LM have very different ranges of values. The accumulated probabilities at the end of each partial sequence of words Ω in the Viterbi trellis is a combination of acoustic P(A/Ω) and language P(Ω) probabilities. Acoustic probabilities are usually smaller than language probabilities and are applied many more times. The gap among accumulated probabilities is therefore usually bigger than the gap among LM probabilities. The immediate consequence is that LM probabilities are irrelevant in most situations for deciding the best path to choice[1][7]. However, when LM probabilities are raised to a power $\alpha > 1$: $(P(\Omega))^\alpha$, all of them are attenuated, but this attenuation is higher for lower probability values. A bigger gap is therefore obtained between high and low probabilities and then LM probabilities are now more relevant to decide the next word combination. There is a maximum value of α from which LM probabilities are overvalued.

It is important to notice that the smoothing technique clearly defines the LM probability distributions and thus, the "a priori" gap among probabilities. So that, the relationship between the smoothing technique and the exponential scaling factor applied over LM probabilities must be stablished.

3 High and Low Smoothed LMs

The purpose of this work was not to achieve an exhaustive comparison of smoothing techniques like others authors did [1] [2]. The main goal was to observe the relationship between the amount of smoothing given by the smoothing technique and the amount of smoothing achieved by the scaling exponential factor. Thus, two well-known back-off smoothing techniques [8] involving very different amount of discounting have been chosen. Witten-Bell (WBd) and Add-One (AOd) discounting have been used to obtain high and low smoothed LMs respectively. In high-smoothed LMs the probability reserved by the smoothing technique for the unseen events is bigger than in low-smoothed LM. As a consequence the gap be-

[1] This phenomenon is also related to the problem of the negligible impact that transition probabilities have in acoustic models.

tween LM probabilities in high-smoothed LMs is smaller than in low-smoothed models.

The amount of discounting performed by Witten-Bell and Add-one techniques does not need to be adjusted by any additional parameter, like in other well-know techniques, such as Kneser-Ney, linear, etc [1] [2]. In both cases the amount of discounting is fixed and fully defined by the technique.

If $h = (\omega_{i-(n-1)}^{i-1})$ is a history representing a sequence of $n-1$ words, $N(w_i/h)$ is the number of times that word w_i appears after history h, $N(h) = \sum_{\substack{\forall\omega/\\ N(w_i/h)\neq 0}} N(\omega_i/h)$ and $\beta(w_i/h^*)$ is the probability distribution of a more general model (h^* represents a history of words of length less than h), the smoothed LM probability $P(w_i/h)$ is calculated as:

$$P(\omega_i/h) = \begin{cases} (1-\lambda)\frac{N(\omega_i/h)}{N(h)} & N(w_i/h) \neq 0 \\ \Big(\sum_{\substack{\forall w_j/\\ N(w_j/h)\neq 0}} \lambda\frac{N(\omega_j/h)}{N(h)}\Big)\frac{\beta(\omega_i/h^*)}{\sum_{\substack{\forall w_j/\\ N(w_j/h)=0}} \beta(\omega_j/h^*)} & N(w_i/h) = 0 \end{cases} \tag{2}$$

$(1-\lambda)$ represents the discount factor, that is, the amount of probability to be subtracted and then redistributed among unseen events. The discount factor (1-λ) can have very different formulations [1] [2]. In fact, we have given adequate values to (1-λ) to obtain high and low-smoothed LMs using Witten-Bell and Add-One discounting respectively. Those discounting are fully explained in the following paragraphs.

High-Smoothing: Witten-Bell Discounting:
In Witten-Bell, the discount $(1-\lambda)$ depends fundamentally on the number of different events T following the history h. That is:

$$1-\lambda = \frac{N(h)}{N(h)+T} \tag{3}$$

It is widely used since it leads to low text set perplexities when compared to other classical back-off methods [1]. However, a dependence was found [2] between perplexity and the size of the training of the LMs when Witten-Bell discounting was used.
In this case a quite important mass of probability is assigned to unseen events (high-smoothing) and the gap between seen and unseen probabilities is reduced. Combinations of words unseen in training can have a relative high probability in test.

Low-Smoothing: Add-One Discounting:
This is a very simple discounting method, adding one to all the counts. It was calculated as:

$$1-\lambda = \frac{N(h)}{N(h)+1} \tag{4}$$

This method does not usually perform well and thus is not commonly used by itself. Usually it is applied as part of more complicate methods [2] [1].
Since $1 \leq T$, using add-one discounting a smaller mass of probability is redistributed among unseen events (low-smoothing) than using Witten-Bell discounting. The gap among LM probabilities is therefore bigger using Add-One discounting.

3.1 LM Evaluation in Perplexity

Topics related to the obtaining of LMs, such as smoothing techniques, are usually evaluated in terms of perplexity. The test set Perplexity (PP) is based on the mean log probability that a LM assigns to a test set ω_1^L of size L. It is thus based exclusively on the probability of words which actually occur in the test as follows:

$$PP = P(\omega_1^{L)})^{-1/L} = e^{-\frac{1}{L}\sum\limits_{i=1}^{L} log(P(\omega_i/\omega_1^{i-1}))} \tag{5}$$

The test set perplexity measures the branching factor associated to a task, which depends on the number of different words in the text. Low perplexity values are obtained when high probabilities are assigned to the test set events by the LM being evaluated. When the test set includes a high number of unseen combinations of n words, the probability $P(\omega_i/\omega_1^{i-1})$ mainly depends on the smoothing technique. In such a case, $P(\omega_i/\omega_1^{i-1})$ is lower for low-smoothed LMs and, as a consequence bad Perplexity values will be obtained. Thus, high-smoothed techniques lead to good perplexity values when evaluated over test-set including a high number of unseen events. However, this good LM behavior is not always confirmed by the CSR system performance which also includes the acoustic models [4].

4 Experimental Evaluation

In this section the relationship between the two redistributions of the LM probabilities, i.e., the application of the smoothing technique and the scaling factor, is experimentally established. The experimental evaluation was carried out with two Spanish databases of very different levels of difficulty: Bdgeo and Info_Tren.

Bdgeo is a task-oriented Spanish speech corpus [9] consisting of 82000 words and a vocabulary of 1208 words. This corpus represents a set of queries to a Spanish geography database. This is a specific task designed to test integrated systems (acoustic, syntactic and semantic modelling) in automatic speech understanding. The training corpus consisted of 9150 sentences. The test set consisted of 600 sentences. Recording was carried out by 12 speakers in laboratory environments at 16Kz.

[2] This technique is applied in Katz's discounting when all events at one state q are seen more than r times [1]

Info_Tren database has recently been recorded as part of a project to develop a dialogue system. Info_tren is a very difficult task of spontaneous Spanish speech dialogues with a vocabulary of around 2000 words plus 15 different acoustic types of disfluencies such as noises, filled pauses, lengthenings, etc. [10]. The task consisted of 227 Spanish dialogues on train information. They were recorded at 8KHz across telephone lines, applying the well known Wizard of Oz mechanism. The training corpus consisted of 191 dialogues uttered by 63 different speakers (1349 user turns resulting in 16500 words plus 5000 disfluencies). The test set consisted of 36 dialogues corresponding to 12 different speakers (308 user turns including 4000 words plus around 500 disfluencies). Info_tren is the first spontaneous dialog database recorded by Castilian Spanish speakers.

High and low-smoothed n-gram LMs with $n = 2 \dots 4$ were obtained, using Witten-Bell (WBd) and Add-One (AOd) discounting respectively. Table 1 shows the perplexity (PP) results obtained. LMs associated to the Info_Tren database included the disfluencies as part of the vocabulary and there was a quite considerable mismatching between training and test [10]. As a consequence, the Perplexity values associated with this task are quite high.

For both tasks, the best (lowest) PP values were obtained using high-smoothed LMs (WBd). Nevertheless, differences among high and low-smoothed models behavior were more important for Info_Tren task. In this task the number of word sequences appearing in the test set but not appearing in the training set is higher than in Bdgeo task. Higher PP values are obtained using low-smoothed LMs, since they assign lower backoff probabilities than high-smoothed LMs to those sequences. For the Bdgeo task, the best PP values were obtained with 4-grams using both high and low-smoothed LMs. However, for the Info_Tren task the best PP results were reached with 3-grams (trigrams) by high-smoothed LMs and with 2-grams (bigrams) by low-smoothed LM were used.

The LMs in Table 1 were integrated into a Spanish CSR system. 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 were used to represent the acoustic model of the word phonetic chain. Different exponential scaling parameters on LM probabilities were applied (α=$1 \dots 7$). Table 2 shows the CSR performances obtained: the Word Error Rate (WER) and the Average number of Active Nodes (AAN) (including both acoustic and LM nodes) needed to decode a sentence. Optimum performances are emphatised and underlined.

When no scaling factor was applied ($\alpha = 1$) low-smoothed LMs got better performances for both databases. As mentioned above, low-smoothed LMs lead to a bigger gap among LM probabilities than high-smoothed models. Thus, LM probabilities are more significant in the Viterbi trellis and, as a consequence, WER are lower than the obtained when using high-smoothed LMs. Computational cost (ANN) is also lower for low-smoothed LMs because, for a fixed beam-search factor, when differences among probabilities are increased, the number of paths to keep in the lattice are reduced.

Table 1. Perplexity (PP) evaluation of n-grams LMs with $n = 1 \dots 4$ for Bdgeo and Info_Tren tasks. Witten-Bell (WBd), Add-One (AOd) discounting were evaluated.

n	Bdgeo		Info_Tren	
	high smoothing WBd	low smoothing AOd	high smoothing WBd	low smoothing AOd
2	13.1	13.89	36.84	57.22
3	7.53	8.30	34.88	69.87
4	7.17	7.72	36.37	77.33

Table 2. %WER evaluation of n-grams LMs of Table 1 with $n = 2 \dots 4$ for Bdgeo and Info_Tren tasks. Witten-Bell (WBd), Add-One (AOd) discounting were evaluated.

n	α	Bdgeo				Info_Tren			
		high smoothing WBd		low smoothing AOd		high smoothing WBd		low smoothing AOd	
		WER	AAN	WER	AAN	WER	AAN	WER	AAN
n=2	1	41.62	3964	33.29	2209	61.69	3260	56.75	2610
	2	25.80	2588	21.60	1207	50.23	2594	47.15	1827
	3	20.22	1508	17.33	684	43.83	1912	42.13	1199
	4	16.99	764	**14.98**	**416**	41.08	1291	**39.89**	**760**
	5	**15.80**	**380**	15.20	258	**39.60**	**799**	40.75	484
	6	15.95	218	15.93	173	40.32	467	42.30	335
	7	17.01	143	18.14	126	41.75	294	43.92	245
n=3	1	38,85	5189	28,3	2935	58,69	6400	55,60	5172
	2	21,86	2984	16,49	1325	48,72	4668	45,21	3233
	3	15,35	1529	12,5	633	42,14	3172	41,50	1876
	4	11,74	702	**10,98**	**339**	38,72	1978	**39.36**	**1060**
	5	**10,82**	**328**	11,04	193	**38,01**	**1135**	40,24	610
	6	10.85	179	13,08	123	38,41	631	43,13	386
	7	13.04	114	15,67	88	41,58	378	47,41	269
n=4	1	38.50	5374	28.59	3058	58.80	6480	55.20	5380
	2	21.86	3053	16.03	1356	48.90	4720	45.00	3410
	3	14.44	1544	11.91	640	42.25	3286	41.04	2237
	4	10.92	704	10.89	339	38.83	2229	**39.10**	**1250**
	5	10.24	328	**10.67**	**190**	**37.84**	**1269**	41.24	708
	6	**10.22**	**177**	13.44	120	38.63	702	43.70	436
	7	12.48	113	16.46	85	42.31	415	48.26	296

As it was mentioned in Section 2, the gap among LM probabilities is bigger for low-smoothed LMs than for high-smoothed ones. The scaling factor α increases this gap. As a consequence, low-smoothed LMs need lower values of α to get the best CSR performance (see Section 2). In any case, differences between optimum system WER obtained by low and high-smoothing techniques are not very significant.

For Bdgeo task, the best performances were obtained using 4-grams as it was predicted by perplexity. However, for Info_Tren task, optimum performances were also obtained with 4-grams for both low and high smoothed LM in spite of the perplexity predictions. In fact, for Info_Tren task, perplexity increases strongly with n, specially using low-smoothed LMs, but WER decreases with n. The results obtained corroborate that PP is not the most adequate measurement of the smoothing technique.

It has been experimentally established that 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 in many cases is perplexity independent).

5 Concluding Remarks

When smoothed LMs are integrated into the CSR system there are several heuristic parameters that must be taken into account. Due to its great effect on final CSR performances, the exponential scaling factor applied to LM probabilities is one of the most important. This factor increases the gap between LM probabilities to make them more competitive with acoustic probabilities in the Viterbi trellis. In this work, the relationship between the smoothing technique and the scaling factor is established. Low and high smoothed LMs have been evaluated in two Spanish tasks of very different difficulty. Similar optimum CSR performances could obtained applying the adequate value of the scaling factor in each case. Low-smoothed LM reach their optimum CSR performances with lower values of the scaling factor than high smoothed LMs because they have an “a priori” bigger gap among LM probabilities. Experiments showed that an increase of the test set perplexity of a LM does not always mean degradation in the model performance, which depends fundamentally on empirical factors.

References

1. Ney, H., Martin, S., Wessel, F.: Statistical Language Modeling using leaving-one-out. In Young, S., ed.: LM. Kluwer Academic Publishers (1997) 174–207
2. Chen, F.S., Goodman, J.: An empirical study of smoothing techniques for language modeling. Computer, Speech and Language **13** (1999) 359–394
3. Rosenfeld, R.: Two decades of statistical language modeling: Where do we go from here (2000)
4. Clarkson, P., Robinson, R.: Improved language modelling through better language model evaluation measures. Computer, Speech and Language **15** (2001) 39–53
5. Jelinek, F.: Five speculations (and a divertimento) on the themes of h. bourlard, h. hermansky and n. morgan. Speech Communication **18** (1996) 242–246
6. Klakow, D., Peters, J.: Testing the correlation of word error rate and perplexity. Speech Communication **38** (2002) 19–28
7. Varona, A., Torres, 1.: Back-Off smoothing evaluation over syntactic language models. In: Proc. of European Conference on Speech Technology. Volume **3**. (2001) 2135–2138
8. Torres, I., Varona, A.: k-tss language models in a Speech recognition Systems. Computer, Speech and Language **15** (2001) 127–149
9. Díaz, J., Rubio, A., Peinado, A., Segarra, E., Prieto, N., F.Casacuberta: Albayzin: a task-oriented spanish Speech Corpus. In: First Int. Conf. on language resources and evaluation. Volume **11**. (1998) 497–501
10. Rodríguez, L., Torres, I., Varona, A.: Evaluation of sublexical and lexical models of acoustic disfluencies for spontaneous Speech recognition in spanish. In: Proc. of European Conference on Speech Technology. Volume **3**. (2001) 1665–1668