In: *Proceedings of CoNLL-2000 and LLL-2000*, pages 79–82, Lisbon, Portugal, 2000.

Using Perfect Sampling in Parameter Estimation of a Whole Sentence Maximum Entropy Language Model*

F. Amaya[†] and **J. M. Benedí**
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de vera s/n, 46022-Valencia (Spain)
{famaya, jbenedi}@dsic.upv.es

Abstract

The Maximum Entropy principle (ME) is an appropriate framework for combining information of a diverse nature from several sources into the same language model. In order to incorporate long-distance information into the ME framework in a language model, a Whole Sentence Maximum Entropy Language Model (WSME) could be used. Until now MonteCarlo Markov Chains (MCMC) sampling techniques has been used to estimate the paramenters of the WSME model. In this paper, we propose the application of another sampling technique: the Perfect Sampling (PS). The experiment has shown a reduction of 30% in the perplexity of the WSME model over the trigram model and a reduction of 2% over the WSME model trained with MCMC.

1 Introduction

The language modeling problem may be defined as the problem of calculating the probability of a string, $p(w) = p(w_1, \ldots, w_n)$. The probability $p(w)$ is usually calculated via conditional probabilities. The *n-gram* model is one of the most widely used language models. The power of the *n-gram* model resides in its simple formulation and the ease of training. On the other hand, *n-grams* only take into account local information, and important long-distance information contained in the string $w_1 \ldots w_n$ cannot be modeled by it. In an attempt to supplement the local information with long-distance information, hybrid models have been proposed such us (Bellegarda, 1998; Chelba, 1998; Benedí and Sanchez, 2000).

The Maximum Entropy principle is an appropriate framework for combining information of a diverse nature from several sources into the same model: the Maximum Entropy model (ME) (Rosenfeld, 1996). The information is incorporated as *features* which are submitted to constraints. The conditional form of the ME model is:

$$p(y|x) = \frac{1}{Z(x)} e^{\sum_{i=1}^{m} \lambda_i f_i(x,y)} \qquad (1)$$

where λ_i are the parameters to be learned (one for each feature), the f_i are usually characteristic functions which are associated to the features and $Z(x) = \sum_y exp\{\sum_{i=1}^{m} \lambda_i f_i(x,y)\}$ is the normalization constant. The main advantages of ME are its flexibility (local and global information can be included in the model) and its simplicity. The drawbacks are that the paramenter's estimation is computationally expensive, specially the evaluation of the normalization constant $Z(x)$ and that the grammatical information contained in the sentence is poorly encoded in the conditional framework. This is due to the assumption of independence in the conditional events: in the events in the state space, only a part of the information contained in the sentence influences de calculation of the probability (Ristad, 1998).

2 Whole Sentence Maximum Entropy Language Model

An alternative to combining local, long-distance and structural information contained in the sentence, within the maximum entropy framework, is the Whole Sentence Maximum Entropy model (WSME) (Rosenfeld, 1997). The

* This work has been partially supported by the Spanish CYCIT under contract (TIC98/0423-C06).
† Granted by Universidad del Cauca, Popayán (Colombia)

WSME is based in the calculation of unrestricted ME probability $p(w)$ of a whole sentence $w = w_1 \dots w_n$. The probability distribution is the distribution p that has the maximum entropy relative to a prior distribution p_0 (in other words: the distribution that minimize de divergence $D(p||p_0)$) (Della Pietra et al., 1995). The distribution p is given by:

$$p(w) = \frac{1}{Z} p_0(w) e^{\sum_{i=1}^{m} \lambda_i f_i(w)} \qquad (2)$$

where λ_i and f_i are the same as in (1). Z is a (global) normalization constant and p_0 is a prior proposal distribution. The λ_i and Z are unknown and must be learned.

The parameters λ_i may be interpreted as being weights of the features and could be learned using some type of iterative algorithm. We have used the Improved Iterative Scaling algorithm (IIS) (Berger et al., 1996). In each iteration of the IIS, we find a δ_i value such that adding this value to λ_i parameters, we obtain an increase in the the log-likelihood. The δ_i values are obtained as the solution of the m equations:

$$\sum_{w} p(w) f_i(w) e^{\delta_i f^{\#}(w)} - \frac{1}{|\Omega|} \sum_{w \in \Omega} \tilde{p}(w) f_i(w) = 0 \qquad (3)$$

where $i = 1, \dots, m$, $f^{\#}(w) = \sum_{i=1}^{m} f_i(w)$ and Ω is a training corpus. Because the domain of WSME is not restricted to a part of the sentence (context) as in the conditional case, it allows us to combine global structural syntactic information which is contained in the sentence with local and other kinds of long range information such us triggers. Furthermore, the WSME model is easier to train than the conditional one, because in the WSME model we don't need to estimate the normalization constant Z during the training time. In contrast, for each event (x, y) in the training corpus, we have to calculate $Z(x)$ in each iteration of the MEC model.

The main drawbacks of the WSME model are its integration with other modules and the calculation of the expected value in the left part of equation (3), because the event space is huge.

Here we focus on the problem of calculating the expected value in (3). The first sum in (3) is the expected value of $f_i e^{\delta_i f^{\#}}$, and it is obviously not possible to sum over all the sentences. However, we can estimate the mean by using the empirical expected value:

$$E_p\left[f_i e^{\delta_i f^{\#}}\right] \approx \frac{1}{M} \sum_{j=1}^{M} f_i(s_j) e^{\delta_i f^{\#}(s_j)} \qquad (4)$$

where $s_1, \dots, s_M$ is a random sample from $p(w)$. Once the parameters have been learned it is possible to estimate the value of the normalization constant, because $Z = \sum_w e^{\sum_{i=1}^{m} \lambda_i f_i(w)} p_0(w) = E_{p_0}\left[e^{\sum_{i=1}^{m} \lambda_i f_i}\right]$, and it can be estimated by means of the sample mean with respect to p_0 (Chen and Rosenfeld, 1999).

In each iteration of IIS, the calculation of (4) requires sampling from a probability distribution which is partially known (Z is unknown), so the classical sampling techniques are not useful. In the literature, there are some methods like the MonteCarlo Markov Chain methods (MCMC) that generate random samples from $p(w)$ (Sahu, 1997; Tierney, 1994). With the MCMC methods, we can simulate a sample *approximately* from the probability distribution and then use the sample to estimate the desired expected value in (4).

3 Perfect Sampling

In this paper, we propose the application of another sampling technique in the parameter estimation process of the WSME model which was introduced by Propp and Wilson (Propp and Wilson, 1996): the Perfect Sampling (PS). The PS method produces samples from the *exact* limit distribution and, thus, the sampling mean given in (4) is less biased than the one obtained with the MCMC methods. Therefore, we can obtain better estimations of the parameters λ_i.

In PS, we obtain a sample from the limit distribution of an ergodic Markov Chain $X = \{X_n; n \geq 0\}$, taking values in the state space S (in the WSME case, the state space is the set of possible sentences). Because of the ergodicity, if the transition law of X is $P(x, A) := P(X_n \in A | X_{n-1} = x)$, then it has a limit distribution π, that is: if we start a path on the chain in any state at time $n = 0$, then as $n \to \infty$, $X_n \to \pi$. The first algorithm of the family of PS was presented by Propp and Wilson (Propp and Wilson, 1996) under the name Coupling From the Past (CFP) and is as follows: start a path in every state of S at some time $(-T)$ in the past such that at time $n = 0$, all the paths collapse to a unique value (due to the ergodicity). This value is a sample element. In the majority of cases, the state space is huge, so attempting to begin a path in every state is not practical. Thus, we can define a partial stochastic order in the state space and so we only need start two paths: one in the minimum and one in the maximum. The two paths collapse at time $n = 0$ and the value of the coalescence state is a sample element of π. The CFP algorithm first determines the time T to start and then runs the two paths from time $(-T)$ to 0. Information about PS methods may be consulted in (Corcoran and Tweedie, 1998; Propp and Wilson, 1998).

4 Experimental work

In this work, we have made preliminary experiments using PS in the estimation of the expected value (4) during the learning of the parameters of a WSME model. We have implemented the Cai algorithm (Cai, 1999) to obtain perfect samples. The Cai algorithm has the advantage that it doesn't need the definition of the partial order.

The experiments were carried out using a pseudonatural corpus: *"the traveler task"*[1]. The traveler task consists in dialogs between travelers and hotel clerks. The size of the vocabulary is 693 words. The training set has 490,000 sentences and 4,748,690 words. The test set has 10,000 sentences and 97,153 words.

Three kinds of features were used in the WSME model: n-grams (1-grams, 2-grams, 3-grams), distance 2 n-grams (d2-2-grams, d2-3-grams) and triggers. The proposal prior distribution used was a trigram model.

We trained WSME models with different sets of features using the two sampling techniques: MCMC and PS. We measured the perplexity (PP) of each of the models and obtained the percentage of improvement in the PP with respect to a trigram base-line model (see table 1). The first model used MCMC techniques (specifically the Independence Metropolis-Hastings algorithm (IMH) [2]) and features of n-grams and distance 2 n-grams. The second model used a PS algorithm and features of n-grams and distance 2 n-grams. The third model used the IMH algorithm and features of triggers. The fourth used PS and features of triggers. Finally, in order to compare with the classical methods, we included the trigram base-line model.

Method	PP	% Improvement
IMH	3.37115	28
PS	3.46336	26
IMH-T	3.37198	28
PS-T	3.26964	30
Trigram	4.66975	-

Table 1: Test set perplexity of the WSME model over the traveler task corpus: IMH with features of n-grams and d-n-grams (IMH), PS with n-grams and d-n-grams (PS) IMH with triggers (IMH-T), PS with triggers (PS-T). The base-line model is a trigram model (Trigram)

In all cases, the WSME had a better performance than the n-gram model. From the results in Table 1, we see that the use of features of triggers improves the performance of the model more than the use of n-gram features, this may be due to the correlation between the triggers and the n-grams, the n-gram information has been absorbed by the prior distribution and diminishes the effects of the feature of n-grams. We believe this is the reason why PS-T in Table 1 is better than PS. We also see how IMH and IHM-T shows the same improvement, i.e. the use of triggers does not seem improve the perplexity of the model but, this may be due to the sampling technique: the parameter values depends on the estimation of an expected value, and the estimation depends on the sampling. Finally, the PS-T has better perplexity than the IMH-T. The only difference between both of these is the sampling technique,neither of then has the correlation influence in the features, so we think that the improvement may be due to the sampling technique.

5 Conclusion and future works

We have presented a different approach to the sampling step needed in the parameter estimation of a WSME model. Using this technique, we have obtained a reduction of 30% in the perplexity of the WSME model over the base-line

[1]EuTrans ESPRIT-LTR Project 20268

[2]IMH has been reported recently as the most useful MCMC algorithm used in the WSME training process.

trigram model and an improvement of 2% over the model trained with MCMC techniques. We are extending our experiments to a major corpus: the Wall Street Journal corpus and using a set of features which is more general, including features that reflect the global structure of the sentence.
We are working on introducing the grammatical information contained into the sentence to the model; we believe that such information improves the quality of the model significantly.

References

J. R. Bellegarda. 1998. A multispan language modeling framework for large vocabulary speech recognition. *IEEE Transactions on Speech and Audio Processing*, 6 (5):456–467.

J.M. Benedí and J.A. Sanchez. 2000. Combination of n-grams and stochastic context-free grammars for language modeling. *International conference on computational linguistics (COLIN-ACL).*

A.L. Berger, V.J. Della Pietra, and S.A. Della Pietra. 1996. A Maximum Entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72.

H. Cai. 1999. Exact Sampling using auxiliary variables. *Statistical Computing Section, ASA Proceedings.*

C. Chelba. 1998. *A structured Language Model.* PhD Dissertation Proposal, The Johns Hopkins University.

S. Chen and R. Rosenfeld. 1999. Efficient sampling and feature selection in whole sentence maximum entropy language models. *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP).*

J.N. Corcoran and R.L. Tweedie. 1998. Perfect sampling for Independent Metropolis-Hastings chains. preprint. Colorado State University.

S. Della Pietra, V. Della Pietra, and J. Lafferty. 1995. Inducing features of random fields. Technical Report CMU-CS-95-144, Carnegie Mellon University.

J. G. Propp and D. B. Wilson. 1996. Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9:223–252.

J. A. Propp and D. B. Wilson. 1998. Coupling from the Past: User's Guide. *Dimacs series in discrete Mathematics and Theoretical Computer Science*, pages 181–192.

E. S. Ristad, 1998. *Maximum Entropy Modeling Toolkit, Version 1.6 Beta.*

R. Rosenfeld. 1996. A Maximum Entropy approach to adaptive statistical language modeling. *Computer Speech and Language*, 10:187–228.

R. Rosenfeld. 1997. A whole sentence Maximum Entropy language model. *IEEE workshop on Speech Recognition and Understanding.*

S. Sahu. 1997. Bayesian data analysis. Technical report, School of Mathematics, University of Walles.

L. Tierney. 1994. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22:1701–1762.