

# A Whole Sentence Maximum Entropy Language Model

R. Rosenfeld

School of Computer Science  
Carnegie Mellon University  
Pittsburgh, PA 15213

**Abstract** - We introduce a new kind of language model, which models whole sentences or utterances directly using the Maximum Entropy paradigm. The new model is conceptually simpler, and more naturally suited to modeling whole-sentence phenomena, than the conditional ME models proposed to date. By avoiding the chain rule, the model treats each sentence or utterance as a “bag of features”, where features are arbitrary computable properties of the sentence. The model is unnormalizable, but this does not interfere with training (done via sampling) or with use. Using the model is computationally straightforward. The main computational cost of training the model is in generating sample sentences from a Gibbs distribution. Interestingly, this cost has different dependencies, and is potentially lower, than in the comparable conditional ME model.

## 1 Motivation

Conventional statistical language models estimate the probability of a sentence  $s$  by using the chain rule to decompose it into a product of conditional probabilities:

$$\Pr(s) \stackrel{\text{def}}{=} \Pr(w_1 \dots w_n) = \prod_{i=1}^n \Pr(w_i | w_1 \dots w_{i-1}) \stackrel{\text{def}}{=} \prod_{i=1}^n \Pr(w_i | h) \quad (1)$$

where  $h \stackrel{\text{def}}{=} \{w_1, \dots, w_{i-1}\}$  is the *history* when predicting word  $w_i$ . The vast majority of work in statistical language modeling is then devoted to estimating terms of the form  $\Pr(w|h)$ .

The application of the chain rule is technically harmless since it uses an exact equality, not an approximation. However, terms like  $\Pr(w|h)$  may not be the best way to think about estimating  $\Pr(s)$ :

1. Global sentence information such as grammaticality is awkward to encode in a conditional framework.

2. External influences on the sentence (for example, the effect of preceding utterances, or dialog level variables) are equally hard to encode. Furthermore, such influences must be factored into the prediction of every word in the current sentence, causing small but systematic biases in the estimate to be compounded.
3.  $\Pr(w|h)$  is typically approximated based on  $\Pr(w_i|w_{i-k+1}, \dots, w_{i-1})$  (the Markov assumption). Even if such a model is improved by including longer distance information, it still makes many implicit independence assumptions. It is evidenced from looking at language data that these assumptions are often patently false, and that there are significant global dependencies both within- and across sentences.

As a simple example of the limitations of the chain rule approach, consider one aspect of an sentence: its length. In an  $N$ -gram based model, the effect of the number of words in the utterance on its probability cannot be modeled directly. Rather it is an implicit consequence of the  $N$ -gram prediction. It is then corrected by a 'word insertion penalty', the usefulness of which proves that length is an important feature. But the word insertion penalty can only model length as a geometric distribution, which does not fit well with empirical data, especially for short utterances.

As an alternative to the conventional conditional formulation, we propose a Maximum Entropy language model which directly models the distribution of an entire sentence or utterance, and which can be faster to train than Maximum Entropy models presented to date.

## 2 Review of Maximum Entropy Framework

Maximum Entropy (ME, [1]) models are exponential distributions which satisfy given linear constraints. In the last few years, ME has been successfully applied in language modeling to approximate conditional probabilities of the form  $P(w|h)$  [2, 3, 4, 5]; to model prepositional phrase attachment [6]; and to induce features of word spelling [7].

In conditional language modeling, given constraint or "feature" functions  $f_i(h, w)$  and respective desired values, the ME solution has the form:

$$P(w|h) = \frac{1}{Z(h)} \cdot \exp\left[\sum_i \lambda_i f_i(h, w)\right] \quad (2)$$

The parameters  $\lambda_i$  can be derived using the Generalized Iterative Scaling algorithm (GIS, [8]). The computational bottleneck in GIS is in computing the expectations  $E_P[f_i]$  of the feature functions, and especially the normalization constant  $Z(h) = \sum_{w \in \mathcal{V}} \exp[\sum_i \lambda_i f_i(h, w)]$  for each training datapoint  $(h, w)$ . This bottleneck is quite severe: training a model that incorporates  $n$ -gram and trigger features on some 40 million words can take hundreds of workstation days [5, p.210]. It is this bottleneck that hinders the widespread use of the ME framework in language modeling.

In what follows, I propose a ME model and training paradigm which address this computational problem as well as the modeling problem discussed in the section 1.

### 3 Sentence Based ME Models

In the new formulation, the standard ME model (2) is replaced by:

$$P(s) = \frac{1}{Z} \cdot \exp\left(\sum_i \lambda_i f_i(s)\right) \quad (3)$$

The new model does not involve the chain rule and is conceptually simpler. In addition, the normalization constant  $Z$  is global, i.e. a single number  $Z$  is used throughout the model. One way to think about the proposed model is to take the old model and map:  $w \rightarrow s$  and  $h \rightarrow \epsilon$  (the null history).

With the new model, normalization is infeasible, since it involves summation over all possible sentences  $s$ . Similarly, the expectations  $E_P[f_i]$  cannot be computed explicitly. These problems can be overcome as follows.

**During Training:** For bivalued feature functions, we estimate the ratio of  $P(s : f(s)=1)$  to  $P(s : f(s)=0)$ , which is independent of  $Z$ , and from which  $E_P[f_i]$  can be derived.

**During Testing:** Without knowing the normalizing constant  $Z$ , perplexity cannot be computed. However, when interfaced to an application such as a speech recognizer, the value of  $Z$  can be ignored since it is the same for all hypotheses considered for the current sentence.

#### 3.1 Estimating the Expectations

Since explicit summation over all sentences  $s$  is infeasible, we will estimate the expectations  $E_P[f_i]$  by sampling. Gibbs Sampling [9] was used in [7] to sample from the population of character strings. We will now describe how to use it to generate whole sentences from the (unnormalized) distribution  $Q(s) = \exp(\sum_i \lambda_i f_i(s))$ , then present an alternative method which, when applicable, is far more efficient.

**Gibbs Sampling of Sentences:** To generate a single sentence from the distribution characterized by  $Q(s)$ , start from an arbitrary sentence  $s$ , we iterate as follows:

1. Choose a word position  $i$  (randomly or by sweeping across the sentence).
2. For each word  $w$  in the vocabulary  $\mathcal{V}$ , place  $w$  in position  $i$  instead of the current word  $w_i$ , resulting in sentence  $s_w^i$ . Calculate  $Q(s_w^i)$ .
3. Choose a word at random according to the distribution  $\{Q(s_w^i)\}_{w \in \mathcal{V}}$ . Place that word in position  $i$  in the sentence. This constitutes a single step in the random walk in the underlying Markov field.

After enough iterations of the above procedure,<sup>1</sup> the resulting sentence is guaranteed to be an unbiased sample from the Gibbs distribution  $P(s)$ .

**Corrective Sampling:** It is sometimes possible to estimate  $E_P[f_i]$  much more efficiently than with Gibbs sampling. This is when there exists some other distribution, say  $R(s)$ , which is computationally easier to sample from, or for which an unbiased sample already exists. We use the following procedure:

1. Generate sentences from  $R(s)$  (or use the existing sample).
2. For every sentence in the sample, calculate the statistics  $\left[\frac{Q(s)}{R(s)}f_i(s)\right]$  for all  $i$ . Their expectations are  $\langle \frac{Q(s)}{R(s)}f_i(s), R(s) \rangle = E_Q[f_i(s)]$ , as desired.

The sampling function  $R(s)$  must be reasonably close to  $P(s)$ . How well  $R(s)$  approximates  $P(s)$  will determine the size of the sample needed for reliable estimation. One choice for  $R(s)$  is a conventional model which incorporates via linear interpolation the same knowledge sources as the target ME model. Another choice for  $R(s)$  is the exponential distribution  $Q(s)$  from a previous iteration of the GIS algorithm, for which a sample already exists.

### 3.2 Feature Types

The most straightforward feature or constraint types to be implemented are those used in conventional  $N$ -gram models or those that have been used in conditional ME models to date: ngrams, distance- $d$  ngrams, class-based ngrams, and word triggers. See [5] for an implementation of these feature types in a conditional ME model.

But, hopefully, this is just the beginning. It is our hope that a Maximum Entropy model of the type discussed above will break the ME “usability barrier” which currently hinders exploration and integration of multiple knowledge sources. If successful, this will open the floodgates to experimentation, by many researchers, with varied knowledge sources which they believe to carry significant information. Such sources may include:

- Sentence length.
- Number and type of verbs present.
- Various aspects of grammaticality (person agreement, number agreement, (partial) parsability, other parser-supplied information).
- Dialog level information.
- Prosodic and other time related information (speaking rate, pauses, . . .).

---

<sup>1</sup>The exact value of “enough” is not theoretically known.

**Filler features:** It is sometimes useful to define a feature to complement a set of existing features. For example, the set  $B(w_1)$  of bigram features whose first word is  $w_1$  may be supplemented by introducing a feature which is “on” only when  $w_1$  is present and none of the features in  $B(w_1)$  are “on”. Similar “filler features” can be defined for other feature types. They can also be defined based on word position and sentence length, which helps in direct modeling of the latter.

Filler features contribute to the numerical stability of the training procedure. Furthermore, they enable the model to assign reasonable probabilities to sentences where some of the feature values are not well defined or else not known.

## 4 Validation

To validate the approach discussed above, we built a whole-utterance ME model using a small (10KW) training set of broadcast news utterances<sup>2</sup>. We used unigram, bigram and trigram features, and appropriate filler features. The features were not introduced all at the same time. Instead, the unigram features were introduced first, and the model was allowed to converge. Next the bigram features were introduced, and the model again allowed to converge. Finally the trigram features were introduced, and the model again allowed to converge. This resulted in faster convergence than in simultaneous introduction of all feature types.

Since the model is unnormalized<sup>3</sup>, it is impossible to compare its perplexity to that of a comparable conventional model. Instead, we provide sample sentences generated by Gibbs sampling from various stages of the training procedure. Table 1 lists sample sentences generated by the initial model, before any training took place. Since the initial  $\lambda$ 's were all set to zero, the resulting model is the uniform model. Tables 2 through 4 list sample sentences generated by the converged model after the introduction of unigram, bigram and trigram features, respectively. It can be seen from the example sentences that the model indeed successfully incorporated the information provided by the respective features.

Although the model described above incorporates only “conventional” features which are easy to incorporate in a simple relative frequency model, this was done for the purpose of supporting future comparisons. The model is unaware of the nature of the features. Arbitrary features can be accommodated with virtually no change in the model or the code.

---

<sup>2</sup>Throughout this paper we have been referring to the unit of modeling as a “sentence”. But of course the method can be used to model any word sequence or utterance, whether or not it is consistent with linguistic boundaries. Naturally, linguistically induced features may or may not be applicable to non-sentences.

<sup>3</sup>Indeed, unnormalizable. It is possible to prove that the normalization constant  $Z$  cannot be feasibly computed.

<s> EVIDENCE COULD QUESTION WE'LL ALSO DO GETTING THEN </s>  
<s> ENOUGH CARE GREG GETTING IF O. ANSWER NEVER </s>  
<s> DEATH YOU'VE BOTH THEM RIGHT BACK WELL BOTH </s>  
<s> MONTH THAT'S NEWS ANY YOU'VE WROTE MUCH MEAN </s>  
<s> A. HONOR WE'VE ME GREG LOOK TODAY N. </s>

Table 1: Sentences generated by Gibbs sampling from the initial (untrained) model. Since all  $\lambda$ 's were initialized to zero, this is the uniform model.

<s> WELL VERY DON'T A ARE NOT LIVE THE </s>  
<s> I RIGHT OF NOT SO ANGELES IS DONE </s>  
<s> I ARE FOUR THIS KNOW DON'T ABOUT OF </s>  
<s> C. GO ARE TO A IT HAD SO </s>  
<s> OFF THE MORE JUST POINT WANT MADE WELL </s>

Table 2: Sentences generated by Gibbs sampling from a whole-utterance ME model trained on unigram features only.

## 5 Summary and Analysis

We have introduced a new kind of language model, which incorporates arbitrary knowledge sources using the Maximum Entropy paradigm, but which is more naturally suited to modeling whole-sentence phenomena than the conditional ME models proposed to date. By avoiding the chain rule, the model treats each sentence or utterance as a “bag of features”, where features are arbitrary computable properties of the sentence. The underlying equations are considerably simpler than those of conditional ME models.

### 5.1 Computational Considerations

Using the model is straightforward and not computationally expensive. The main computational burden of training the model is in generating the sample sentences. Indeed, efficient sampling is crucial for practical success.

<s> DO YOU WANT TO DON'T C. WAS YOU </s>  
<s> THE I DO YOU HAVE A A US </s>  
<s> BUT A LOS ANGELES ASK C. NEWS ARE </s>  
<s> WE WILL YOU HAVE TO BE AGENDA AND </s>  
<s> THE WAY IS THE DO YOU THINK ON </s>

Table 3: Adding bigram features.

```
<s> WHAT DO YOU HAVE TO LIVE LOS ANGELES </s>  
<s> A. B. C. N. N. BUSINESS NEWS TOKYO </s>  
<s> BE OF SAYS I'M NOT AT THIS IT </s>  
<s> BILL DORMAN BEEN WELL I THINK THE MOST </s>  
<s> DO YOU HAVE TO BE IN THE WAY </s>
```

Table 4: Adding trigram features.

Interestingly, the computational requirements have a different type of dependence than the comparable conditional ME models. Specifically, the computational cost is determined mostly by the combination of how *rare* the features are and how *accurately* we want to model them.

Much experimental work still needs to be done to improve the efficiency of the sampling process. In particular the following heuristics should be explored:

- At each iteration, consider only a subset of the vocabulary for replacement. This trades off the computational cost per iteration against the “mixing rate” of the underlying Markov chain.
- Use only rough estimation in the first few iterations (we only need to know the direction and rough magnitude of the correction to the  $\lambda$ s); increase sample size when approaching convergence.
- Determine the sample size dynamically.

## 5.2 Modeling Considerations

A whole-sentence ME model incorporating the same features as a conditional ME model is in fact not identical to the latter. This is because the training procedure used for conditional ME models restricts the computation of the feature expectations to histories observed in the training data (see [4] or [5, section 4.4]). This biases the solution in an interesting and usually appropriate way. For example, consider two word trigger features:  $A \rightarrow Z$  and  $B \rightarrow Z$ . If  $A$  and  $B$  are correlated in the training data, this will affect the solution of the conditional ME model. In fact, if they are perfectly correlated, always co-occurring in the the same document, the resulting  $\lambda$ s will be one half of what their value would be if only one of the features was used. This is beneficial to the model, since it captures correlations that are likely to recur in new data. However, a whole-sentence ME model incorporating the same features will not use the correlation between  $A$  and  $B$ , unless it is explicitly instructed to do so via a separate feature (the training data is not actually used in whole-sentence ME training, except to derive the target values).

## Acknowledgements

I am grateful to Sanjeev Khudanpur, Fred Jelinek and Prakash Narayan for helpful discussions and suggestions. This project was sponsored by the National Security Agency under Grant number MDA904-97-1-0006. The United States Government is authorized to reproduce and distribute reprints notwithstanding any copyright notation herein.

## References

- [1] E. T. Jaynes, "Information Theory and Statistical Mechanics," *Physics Reviews*, vol. 106, pp. 620–630, 1957.
- [2] S. Della Pietra, V. Della Pietra, R. Mercer and S. Roukos, "Adaptive Language Modeling Using Minimum Discriminant Estimation," In *Proceedings ICASSP*, 1992, pp. I-633–636.
- [3] R. Lau, R. Rosenfeld and S. Roukos, "Trigger-Based Language Models: a Maximum Entropy Approach," In *Proc. ICASSP*, 1993, pp. II 45–48.
- [4] A. Berger, S. Della Pietra and V. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, vol. 22, no. 1, March 1996.
- [5] R. Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling," *Computer, Speech and Language*, vol. 10, pp. 187–228, 1996. Longer version: *Carnegie Mellon Tech. Rep. CMU-CS-94-138*.
- [6] A. Ratnaparkhi and S. Roukos, "A Maximum Entropy Model for Prepositional Phrase Attachment," In *Proceedings of the ARPA Workshop on Human Language Technology*, Morgan Kaufmann, 1994.
- [7] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing features of random fields," In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 380–393, April 1997.
- [8] J. N. Darroch and D. Ratcliff, "Generalized Iterative Scaling for Log-Linear Models," *The Annals of Mathematical Statistics*, vol. 43, pp. 1470–1480, 1972.
- [9] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.