SMOOTHING METHODS IN MAXIMUM ENTROPY LANGUAGE MODELING

S. C. Martin, H. Ney, J. Zaplo

Lehrstuhl für Informatik VI, RWTH Aachen, University of Technology, D-52056 Aachen, Germany

ABSTRACT

This paper discusses various aspects of smoothing techniques in maximum entropy language modeling, a topic not sufficiently covered by previous publications. We show (1) that straightforward maximum entropy models with nested features, e.g. tri–, bi–, and unigrams, result in unsmoothed relative frequencies models; (2) that maximum entropy models with nested features and discounted feature counts approximate backing–off smoothed relative frequencies models with Kneser's advanced marginal back–off distribution; this explains some of the reported success of maximum entropy models in the past; (3) perplexity results for nested and non–nested features, e.g. trigrams and distance–trigrams, on a 4–million word subset of the Wall Street Journal Corpus, showing that the smoothing method has more effect on the perplexity than the method to combine information.

1. MAXIMUM ENTROPY APPROACH

The maximum entropy principle [1, 5] is a well-defined method for incorporating different types of features into a language model [4, 9]. For a word w given its history h it has the following functional form [2, pp. 83-87]:

$$p_\Lambda(w|h) = \frac{\exp\left[\sum_i \lambda_i f_i(h,w)\right]}{Z(h)} \qquad (1)$$

$$Z(h) := \sum_{\tilde{w}} \exp\left[\sum_i \lambda_i f_i(h,\tilde{w})\right] \quad ,$$

where for each feature i we have a feature function $f_i(h,w) \in \{0,1\}$ that is activated if feature i exists in (h,w), and a weight parameter λ_i, with parameter set $\Lambda := \{\lambda_i\}$. Considering *conditional* probabilities is an important difference to most standard publications. For the parameter estimation of Λ we consider the log-likelihood function $G(\Lambda)$ for a training corpus of running words $w_1, ..., w_n, ..., w_N$:

$$G(\Lambda) := \sum_{n=1}^{N} \log p_\Lambda(w_n|h_n) = \sum_{hw} N(h,w) \log p_\Lambda(w|h) \quad ,$$

with the usual count definitions $N(h,w)$. We take the partial derivatives with respect to each of the parameters λ_i, set them to zero, and obtain the so–called constraint equation for each feature i:

$$\frac{\partial G}{\partial \lambda_i} = -Q_i(\Lambda) + N_i = 0$$

$$Q_i(\Lambda) := \sum_{hw} N(h)\, p_\Lambda(w|h)\, f_i(h,w)$$

$$N_i := \sum_{hw} N(h,w)\, f_i(h,w) \quad ,$$

with the Λ–dependent auxiliary function $Q_i(\Lambda)$ and the Λ–independent feature counts N_i. There is no closed solution to the set of constraint equations. We train them with the Generalized Iterative Scaling (GIS) algorithm [3] implemented as described in [10] with the addition of Ristad's speedup technique [11].

In this paper the baseline maximum entropy model uses the nested trigram, bigram, and unigram features with $(h,w) = (u,v,w)$:

$$f_{uvw}(\tilde{u},\tilde{v},\tilde{w}) = \begin{cases} 1 & \text{if } w=\tilde{w} \text{ and } v=\tilde{v} \text{ and } u=\tilde{u} \\ 0 & \text{otherwise} \end{cases}$$

$$f_{vw}(\tilde{u},\tilde{v},\tilde{w}) = \begin{cases} 1 & \text{if } w=\tilde{w} \text{ and } v=\tilde{v} \\ 0 & \text{otherwise} \end{cases}$$

$$f_{w}(\tilde{u},\tilde{v},\tilde{w}) = \begin{cases} 1 & \text{if } w=\tilde{w} \\ 0 & \text{otherwise} \end{cases}$$

Motivated by the good results in [10], the baseline model is extended by non–nested features, either distance–2–trigrams with $(h,w) = (t,u,v,w)$:

$$f_{t\cdot vw}(\tilde{t},\tilde{u},\tilde{v},\tilde{w}) = \begin{cases} 1 & \text{if } w=\tilde{w} \text{ and } v=\tilde{v} \text{ and } t=\tilde{t} \\ 0 & \text{otherwise} \end{cases}$$

$$f_{tu\cdot w}(\tilde{t},\tilde{u},\tilde{v},\tilde{w}) = \begin{cases} 1 & \text{if } w=\tilde{w} \text{ and } u=\tilde{u} \text{ and } t=\tilde{t} \\ 0 & \text{otherwise} \end{cases}$$

or, alternatively, distance–3–bigrams and distance–4–bigrams with $(h,w) = (s,t,u,v,w)$:

$$f_{t\cdot\cdot w}(\tilde{s},\tilde{t},\tilde{u},\tilde{v},\tilde{w}) = \begin{cases} 1 & \text{if } w=\tilde{w} \text{ and } t=\tilde{t} \\ 0 & \text{otherwise} \end{cases}$$

$$f_{s\cdots w}(\tilde{s},\tilde{t},\tilde{u},\tilde{v},\tilde{w}) = \begin{cases} 1 & \text{if } w=\tilde{w} \text{ and } s=\tilde{s} \\ 0 & \text{otherwise} \end{cases}$$

As opposed to the non–nested features, closed solutions exist in part for nested features, allowing some analysis of the maximum entropy models.

2. SMOOTHING OF MAXIMUM ENTROPY MODELS

2.1. Unsmoothed Models: Relative Frequencies

For the straightforward baseline maximum entropy model, there is a closed solution due to the nested features. The constraint equations

$$\sum_{\tilde{u},\tilde{v},\tilde{w}} N(\tilde{u},\tilde{v}) \cdot p_\Lambda(\tilde{w}|\tilde{u},\tilde{v}) \cdot f_{uvw}(\tilde{u},\tilde{v},\tilde{w})$$

$$= N(u,v) \cdot \frac{e^{\lambda_{uvw}+\lambda_{vw}+\lambda_w}}{Z(u,v)} = N(u,v,w)$$

result in relative frequencies:

$$\frac{e^{\lambda_{uvw}+\lambda_{vw}+\lambda_w}}{Z(u,v)} = \frac{N(u,v,w)}{N(u,v)} \quad .$$

Since the probabilities of all seen trigrams of a given history (u, v) sum up to 1, the probability of unseen trigrams is not properly defined using the model of Eq. (1) because of $e^{\lambda_{uvw}+\lambda_{vw}+\lambda_w} > 0$, even though bigram and unigram features exist for backing–off. Therefore, smoothing must be applied, a technique that redistributes probability mass from seen to unseen events [8].

2.2. Smoothing Using Cut–Offs and Absolute Discounting

We do not know an obvious smoothing technique for maximum entropy, so we adapted techniques from known smoothing methods:

- **Cut–Offs:** Probability mass is gained by omitting features i with a feature count N_i of threshold k and below. However, this results in a coarser model.
- **Absolute Discounting:** This method was first presented in [10] without detailed analysis. All features i with a positive feature count are allowed. Probability mass is gained by reducing the feature count N_i by a fixed discounting value d. It is important to note that we now diverge from the maximum likelihood principle and risk inconsistent constraint equations. In the experiments, we use three different discounting values d_T, d_B, and d_U for trigram, bigram and unigram features, respectively.

We analyse the effect of the smoothing methods for the case that all bigram features are seen and thus not smoothed. This is unrealistic but leads to a closed solution. If we apply both smoothing methods at the same time, we get the model:

$$p_\Lambda(w|u,v) = \begin{cases} \dfrac{e^{\lambda_{uvw}+\lambda_{vw}}}{Z(u,v)} & \text{if } N(u,v,w) > k \\ \dfrac{e^{\lambda_{vw}}}{Z(u,v)} & \text{otherwise} \end{cases} \tag{2}$$

and the constraint equations for the upper branch:

$$\sum_{\tilde{u},\tilde{v},\tilde{w}} N(\tilde{u},\tilde{v}) \cdot p_\Lambda(\tilde{w}|\tilde{u},\tilde{v}) \cdot f_{uvw}(\tilde{u},\tilde{v},\tilde{w})$$
$$= N(u,v) \cdot \frac{e^{\lambda_{uvw}+\lambda_{vw}}}{Z(u,v)} \overset{!}{=} N(u,v,w) - d \quad ,$$

resulting in:

$$\frac{e^{\lambda_{uvw}+\lambda_{vw}}}{Z(u,v)} = \frac{N(u,v,w)-d}{N(u,v)} \quad . \tag{3}$$

Eq. (3) is used for the solution of the constraint equations for the lower branch:

$$\sum_{\tilde{u},\tilde{v},\tilde{w}} N(\tilde{u},\tilde{v}) \cdot p_\Lambda(\tilde{w}|\tilde{u},\tilde{v}) \cdot f_{vw}(\tilde{u},\tilde{v},\tilde{w})$$
$$= \sum_{\tilde{u}:\ N(\tilde{u},v,w)\le k} N(\tilde{u},v) \cdot \frac{e^{\lambda_{vw}}}{Z(\tilde{u},v)}$$
$$+ \sum_{\tilde{u}:\ N(\tilde{u},v,w)>k} N(\tilde{u},v,w) - d = N(v,w) \quad ,$$

resulting in:

$$\begin{aligned} e^{\lambda_{vw}} &= \frac{N_{\le k}(\cdot,v,w) + n_{>k}(\cdot,v,w)\cdot d}{\displaystyle\sum_{\tilde{u}:\ N(\tilde{u},v,w)\le k} \frac{N(\tilde{u},v)}{Z(\tilde{u},v)}} \\ &\approx \frac{N_{\le k}(\cdot,v,w) + n_{>k}(\cdot,v,w)\cdot d}{\displaystyle\sum_{\tilde{u}} \frac{N(\tilde{u},v)}{Z(\tilde{u},v)}} \quad , \end{aligned} \tag{4}$$

with

$$\begin{aligned} N_{\le k}(\cdot,v,w) &:= \sum_{\tilde{u}:\ N(\tilde{u},v,w)\le k} N(\tilde{u},v,w) \\ n_{>k}(\cdot,v,w) &:= \sum_{\tilde{u}:\ N(\tilde{u},v,w)>k} 1 \quad . \end{aligned}$$

The approximation is possible because almost all trigrams are unseen in real cases. The $Z(u,v)$ computation, also using Eq. (3), results in

$$Z(u,v) = \frac{\displaystyle\sum_{\tilde{w}:\ N(u,v,\tilde{w})\le k} e^{\lambda_{v\tilde{w}}}}{\dfrac{N_{\le k}(u,v,\cdot) + n_{>k}(u,v,\cdot)\cdot d}{N(u,v)}} \quad , \tag{5}$$

with

$$\begin{aligned} N_{\le k}(u,v,\cdot) &:= \sum_{\tilde{w}:\ N(u,v,w)\le k} N(u,v,\tilde{w}) \\ n_{>k}(u,v,\cdot) &:= \sum_{\tilde{w}:\ N(u,v,\tilde{w})>k} 1 \quad . \end{aligned}$$

Note that
$$\frac{N_{\le k}(u,v,\cdot) + n_{>k}(u,v,\cdot)\cdot d}{N(u,v)}$$
is the probability mass for redistribution for $h = (u,v)$. Inserting Eqs. (3), (4), (5) into the original model Eq. (2) results in the final backing–off model. For cut–offs ($d = 0$, $k > 0$) we have:

$$p_\Lambda(w|u,v) =$$
$$= \begin{cases} \dfrac{N(u,v,w)}{N(u,v)} & \text{if } N(u,v,w) > k \\ \dfrac{N_{\le k}(u,v,\cdot)}{N(u,v)} \cdot \beta_{uv}(w) & \text{otherwise} \end{cases}$$

with

$$\beta_{uv}(w) := \frac{N_{\le k}(\cdot,v,w)}{\sum_{\tilde{w}:\ N(u,v,w)\le k} N_{\le k}(\cdot,v,\tilde{w})} \quad .$$

Thus, the resulting model is a standard backing–off model [8], but with a back–off distribution $\beta_{uv}(w)$ not known from previous publications. However, this back–off distribution is not properly defined if $\sum_{\tilde{w}:\ N(u,v,w)\le k} N_{\le k}(\cdot,v,\tilde{w}) = 0$. For absolute discounting ($0 < d < 1$, $k = 0$), we have

$$p_\Lambda(w|u,v) =$$
$$= \begin{cases} \dfrac{N(u,v,w)-d}{N(u,v)} & \text{if } N(u,v,w) > 0 \\ \dfrac{n_{>0}(u,v,\cdot)\cdot d}{N(u,v)} \cdot \beta_{uv}(w) & \text{otherwise} \end{cases}$$

Table 1: CPU hours per GIS iteration for maximum entropy models with different features on an Alpha 21164 500 Mhz processor on the WSJ0–4M corpus, $k = 0$.

Features	CPU hours
tri–/bi–/unigrams	2.8
tri–/bi–/unigrams, distance–2–trigrams	3.8
tri–/bi–/unigrams, distance–3/4–bigrams	10.3

with

$$\beta_{uv}(w) := \frac{n_{>0}(\cdot, v, w)}{\sum_{\tilde{w}:\ N(u,v,\tilde{w})=0} n_{>0}(\cdot, v, \tilde{w})} \quad .$$

Thus, the resulting model is a standard backing–off model [8] with a back–off distribution $\beta_{uv}(w)$ known as Kneser's marginal distribution [7]. A closed solution including unigram features is not yet found for both smoothing approaches, but we assume that the resulting models would be similar to the above.

3. EXPERIMENTAL RESULTS

For the experiments, a 4.5–million word text from the Wall Street Journal task was used (exact size: 4,472,827 words). The vocabulary consisted of approximately 20,000 words (`vocab20o.nvp`). All other words in the text were replaced by the label `<UNK>` for the unknown word. The test set perplexity was calculated on a separate text of 325,000 words. In the perplexity calculations, the unknown word was included. The corpora used are the same as in [8] and [9]. The CPU time needed for the improved GIS training can be seen in Table 1.

For nested features, we compared the two smoothing methods for maximum entropy with known smoothing methods for relative frequencies [8]: (1) backing–off with absolute discounting and relative frequencies back–off distribution $\beta(\cdot)$, (2) the same, but with Kneser's marginal back–off distribution, and (3) the standard smoothing method at our site, interpolation with absolute discounting and singleton back–off distribution:

$$\begin{aligned} q(w|u,v) &= \frac{\max\{N(u,v,w)-d, 0\}}{N(u,v)} \\ &\quad + \frac{n_{>0}(u,v,\cdot)\cdot d}{N(u,v)} \cdot \beta(w|v) \\ \beta(w|v) &= \frac{n_1(\cdot,v,w)}{\sum_{\tilde{w}} n_1(\cdot,v,\tilde{w})} \end{aligned}$$

with

$$n_1(\cdot,v,w) := \sum_{\tilde{u}:\ N(\tilde{u},v,w)=1} 1 \quad .$$

The respective smoothing methods were recursively applied to the back–off distributions $\beta(\cdot)$. For smoothed relative frequencies and maximum entropy models the discounting parameters d were estimated by leaving–one–out [8]. In Table 2 we see that the perplexity for the maximum entropy model with absolute discounting is better than explicit backing–off with relative frequencies as back–off distributions, but worse than explicit backing–off with marginal back–off distributions. This underlines that the discounted maximum entropy model only approximates the latter. The cut–off maximum entropy model performs worse, probably because of the poorer modeling and the problematic back–off distribution $\beta(\cdot)$. The standard model performs best, because it employs interpolation instead of backing–off for smoothing. The superiority of smoothing by interpolation over smoothing by backing–off has been observed earlier [8]. Interpolating the standard model with the maximum entropy models

$$p(w|u,v) = (1-\mu)\cdot q(w|u,v) + \mu \cdot p_\Lambda(w|u,v)$$

results in a modest improvement only. All these results show that the performance of a language model with nested features is clearly dominated by the smoothing method, not by the way the features are combined. A baseline maximum entropy model with a better smoothing method or more efficient features may exist but still has to be found.

Table 2: Test set perplexities for trigram language models with different smoothing methods on the WSJ0–4M corpus.

Model	PP
smoothed relative frequencies:	
backing–off	163.4
backing–off, marginal distribution	153.2
standard model	152.1
maximum entropy (20 iterations):	
cut–offs, $k = 1$	178.7
absolute discounting	157.9
interpolation of standard model and maximum entropy (20 iterations):	
cut–offs, $k = 4$, $\mu = 0.3$	148.2
absolute discounting, $\mu = 0.3$	150.0

For non–nested features we compared the effects of extending the models by distance–2–trigrams. For smoothed relative frequencies, each of the three trigram models was separately smoothed by absolute discounting and interpolation, like the standard model, with and without the singleton back–off distribution $\beta(\cdot)$. The discounting parameters d were estimated using leaving–one–out. The three smoothed models were combined by linear interpolation

$$\begin{aligned} &p(w|s,t,u,v) = \\ &= (1-\mu_1-\mu_2)\cdot q(w|u,v) + \mu_1 \cdot q(w|t,\cdot,v) + \mu_2 \cdot q(w|t,u,\cdot) \end{aligned}$$

with interpolation parameters μ_1, μ_2 estimated by a simplified cross validation method. The contesting maximum entropy model was extended by the distance–2–trigram features and initialized for GIS training with the parameters from the baseline nested trigram model. The discounting parameter d_D for absolute discounting for both distance–2–trigram features and the number of GIS iterations was optimized on the testing data. Thus, the training procedure was slightly in favour of the maximum entropy models. Even though, as seen from Table 3, the maximum entropy models are still outperformed by the smoothed relative frequencies model with marginal back–off distribution. The interpolation of the maximum entropy model with the standard model results in a slight perplexity improvement only. Again, results are dominated by the smoothing method.

Table 3: Test set perplexities for trigram and distance–2–trigram language models on the WSJ0–4M corpus.

Model	PP
interpolation of smoothed relative frequencies:	
without singleton smoothing	151.5
with singleton smoothing	138.6
maximum entropy:	
absolute discounting, $d_D = 0.5$, 3 iter.	146.9
interpolation of standard model and maximum entropy:	
absolute discounting, $d_D = 0.5$, 3 iter., $\mu = 0.5$	141.9

Table 4: Test set perplexities for trigram, distance–3–bigram and distance–4–bigram language models on the WSJ0–4M corpus.

Model	PP
interpolation of smoothed relative frequencies:	
singleton distribution	148.6
maximum entropy:	
absolute discounting, $d_D = 0.5$, 3 iter.	147.1
interpolation of standard model and maximum entropy:	
absolute discounting, $d_D = 0.5$, 3 iter., $\mu = 0.6$	142.7

The extension of the trigram models by distance bigrams was performed in the very same way, but with a slightly different result. As can be seen from Table 4, the maximum entropy model now reaches the performance of the smoothed relative frequencies model. An explanation could be that smoothing has a weaker effect on bigrams because bigrams are better trained than trigrams. Thus, the way in which the features are combined becomes more dominant, obviously in favour of the maximum entropy model, as theory suggests [1, 9].

Compared to the backing–off smoothed relative frequencies model without marginal back–off distribution we get a reduction in perplexity by 10% for the maximum entropy model with distance–m–gram features. A similar figure is reported in [9] using Turing–Good smoothing [6] for the maximum entropy model [9, p. 204], a smoothing method comparable to absolute discounting [8]. However, as can be seen from Table 2, roughly a third of this perplexity reduction is already achieved by the marginal back–off distribution implicitly modeled by the maximum entropy model without distance–m–grams, a fact not discussed in earlier publications.

4. CONCLUSION

In this paper we discussed various aspects of smoothing techniques in maximum entropy language modeling. For nested features,

- the unsmoothed maximum entropy model leads to relative frequencies without proper probabilities for events not seen in the training;
- discounted feature counts approximate the well–known backing–off smoothing implicitly using Kneser's advanced marginal back–off distribution;
- the discounted maximum entropy model is outperformed by relative frequencies models with state–of–the–art smoothing.

For non–nested features,

- no closed solutions are known;
- if smoothing is imortant, smoothing methods, not the method of integrating information, dominate the global performance of language models;
- if the features become better trained, smoothing becomes less important, and maximum entropy appears to outperform linear interpolation.

The authors would like to thank Christoph Hamacher for his support in the experiments.

5. REFERENCES

[1] A. L. Berger, S. Della Pietra, V. Della Pietra: "A Maximum Entropy Approach to Natural Language Processing", *Computational Linguistics*, Vol. 22, No. 1, pp. 39–71, 1996.

[2] Y. M. M. Bishop, S. E. Fienberg, P. W. Holland: *Discrete Multivariate Analysis*, MIT press, Cambridge, MA, 1975.

[3] J. N. Darroch, D. Ratcliff: "Generalized Iterative Scaling for Log–Linear Models", *Annals of Mathematical Statistics*, Vol. 43, pp. 1470–1480, 1972.

[4] S. Della Pietra, V. Della Pietra, R. L. Mercer, S. Roukos: "Adaptive Language Modeling Using Minimum Discriminant Information", *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, San Francisco, CA, Vol. I, pp. 633–636, 1992.

[5] S. Della Pietra, V. Della Pietra, J. Lafferty: "Inducing Features of Random Fields", *Technical Report* CMU-CS-95-144, Carnegie Mellon University, Pittsburgh, PA, 1995.

[6] I. J. Good: "The Population Frequencies of Species and the Estimation of Population Parameters", *Biometrika*, Vol. 40, pp. 237-264, Dec. 1953.

[7] R. Kneser, H. Ney: "Improved Backing-Off for m–gram Language Modeling", *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Detroit, MI, Vol. I, pp. 181–184, May 1995.

[8] H. Ney, F. Wessel, S. C. Martin: "Statistical Language Modeling Using Leaving-One-Out", In S. Young, G. Bloothooft (eds.): *Corpus-Based Methods in Speech and Language*, Kluwer Academic Publishers, pp. 174-207, 1997.

[9] R. Rosenfeld: "A Maximum Entropy Approach to Adaptive Statistical Language Modeling", *Computer Speech and Language*, Vol. 10, No. 3, pp. 187–228, July 1996.

[10] M. Simons, H. Ney, S. C. Martin: "Distant Bigram Language Modelling using Maximum Entropy", *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Munich, Vol. II, pp. 787–790, April 1997.

[11] A. Stolcke, C. Chelba, D. Engle, V. Jimenez, L. Mangu, H. Printz, E. Ristad, R. Rosenfeld, D. Wu, F. Jelinek, S. Khudanpur: "Dependency Language Modeling", *1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Reports*, Research Note 24, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, April 1997.