MAXIMUM ENTROPY MODEL PARAMETERIZATION WITH TF*IDF WEIGHTED VECTOR SPACE MODEL

Ye-Yi Wang and Alex Acero

Microsoft Research

ABSTRACT

Maximum entropy (MaxEnt) models have been used in many spoken language tasks. The training of a MaxEnt model often involves an iterative procedure that starts from an initial parameterization and gradually updates it towards the optimum. Due to the convexity of its objective function (hence a global optimum on a training set), little attention has been paid to model initialization in MaxEnt training. However, MaxEnt model training often ends early before convergence to the global optimum, and prior distributions with hyper-parameters are often added to the objective function to prevent over-fitting. This paper shows that the initialization and regularization hyper-parameter setting may significantly affect the test set accuracy. It investigates the MaxEnt initialization/regularization based on an n-gram classifier and a TF*IDF weighted vector space model. The theoretically motivated TF*IDF initialization/regularization has achieved significant improvements over the baseline flat initialization/regularization, especially when training data are sparse. In contrast, the n-gram based initialization/ regularization does not exhibit significant improvements.

Index Terms **—** Maximum entropy model, TF*IDF, vector space model, n-gram classification model, model initialization, model regularization.

1. INTRODUCTION

Maximum entropy (MaxEnt) models have been used in various tasks related to the spoken language technology, including language modeling [1], call-routing [2], and confidence measures [3, 4], etc. The training algorithms for a MaxEnt model, for example, generalized iterative scaling [5] or stochastic gradient ascend [6], involves an iterative procedure that starts from an initial parameterization and gradually updates it towards the optimum. The MaxEnt models have a convex objective function. Hence they converge to a global optimum with respect to a training set. Because of that, little attention has been paid to the initialization of a MaxEnt model. Flat (setting all parameters to 0) or random initialization is commonly used.

However, MaxEnt models are seldom trained to converge at the global optimum with respect to the training data. Early stopping is a common practice to avoid model over-training. A cross-validation set is used during the training procedure to decide when to stop. Therefore, different initializations may end up with different parameterization hence different classification accuracy.

In addition to model initialization, regularization terms are often added to the MaxEnt objective function to restrict the parameters from being astray too much from a specific value. Often a zero-mean Gaussian distribution is used as the prior for parameters (L2-norm regularization). In other words, all features are treated equally a priori. However, if there is an oracle that can tell the importance of different features, the means of the priors should be set accordingly.

This paper investigates the effect of different model initializations and regularization hyper-parameter settings on the accuracy of a MaxEnt model for *text classification*. We have studied the initialization and hyper-parameter setting with the n-gram classification model and the TF*IDF weighted vector space model (Henceforce TF*IDF model). Both models can be trained efficiently without an iterative procedure. Hence they are suitable for model initialization. The TF*IDF model has great success in information retrieval and is difficult to beat. It is discriminative in nature and it is very robust. It provides a weight sharing mechanism for linear classifiers. Researchers have been searching for the theoretical justification for this weighting scheme originally proposed as a heuristic [7]. Recent work has revealed its relation with a relaxed and much simplified MaxEnt model [8].

Preliminary experiments shows that properly scaled TF*IDF initialization/regularization has significantly improved the classification accuracy in different tasks, while the n-gram initialization/regularization for a MaxEnt model does not exhibit significant improvement.

The paper is organized as follows: Section 2 introduces different linear classification models. Section 3 describes the conversion of an n-gram classifier or a TF*IDF model to the parameters and hyper-parameters of a MaxEnt model, and Section 4 presents some preliminary experimental results. Section 5 concludes the paper.

2. TEXT CLASSIFICATION MODELS

2.1. MaxEnt Model

A MaxEnt classifier models the conditional probability distribution $\mathrm{P}(C \mid Q)$ from a set of features $\mathcal{F}$, where C is a random variable representing the classification destinations, Q is a random variable representing input queries. A feature in $\mathcal{F}$ is a function of C and Q. The classifier picks a distribution $\mathrm{P}(C \mid Q)$ to maximize the conditional entropy $H(C \mid Q)$ from a family of distributions, with the constraint that the expected count of a feature predicted by the conditional distribution equals to the empirical count of the feature observed in the training data:

$$\sum_{C,Q} \hat{\mathrm{P}}(Q) \cdot \mathrm{P}(C \mid Q) \cdot f_i(C,Q) = \sum_{C,Q} \hat{\mathrm{P}}(C,Q) \cdot f_i(C,Q), \quad \forall f_i \in \mathcal{F}. \tag{1}$$

where $\hat{\mathrm{P}}$ stands for empirical distributions in a training set.

It has been proven that the maximum entropy distribution that satisfy Eq. (1) have the following exponential (log-linear) form and the parameterization that maximizes the entropy maximizes the conditional probability of a training set of C and Q pairs [9].

$$\mathrm{P}(C \mid Q) = \frac{1}{Z_\lambda(Q)} \exp\left(\sum_{f_i \in \mathcal{F}} \lambda_i f_i(C,Q)\right) \tag{2}$$

$Z_\lambda(Q) = \sum_C \exp(\sum_{f_i \in \mathcal{F}} \lambda_i f_i(C,Q))$ is a normalization constant, and λ_i's are the parameters of the model, also known as the weights of the features. They can be estimated with an iterative procedure that starts from an initial parameterization and gradually updates it towards the optimum. Examples of such training algorithms include Generalized Iterative Scaling (GIS) [5] and Stochastic Gradient Ascend (SGA) [6].

The objective function in (2) is often added with the regularization terms to avoid model over-fitting:

$$L(\lambda) = \frac{1}{Z_\lambda(Q)} \exp\left(\sum_{f_i \in \mathcal{F}} \lambda_i f_i(C,Q)\right) - \sum_i \frac{(\lambda_i - m_i)^2}{2\sigma^2} \tag{3}$$

The regularization terms penalize the parameter λ_i that is too far away from the expected mean value m_i. Without a priori knowledge, m_i is often set to 0.

We applied SGA for model optimization. It is easy to derive the gradient of the objective function as

$$\frac{\partial \log \mathrm{P}(C \mid Q)}{\partial \lambda_i} = \mathsf{E}_{\hat{\mathrm{P}}(Q,C)} f_i(C,Q) - \mathsf{E}_{\hat{\mathrm{P}}(Q)\mathrm{P}(C|Q)} f_i(C,Q) - \frac{\lambda_i - m_i}{\sigma^2} \tag{4}$$

2.2. N-gram Classification Model

An n-gram classifier models the conditional distribution according to a channel model:

$$\mathrm{P}(C \mid Q) \propto \mathrm{P}(C)\mathrm{P}(Q \mid C) = \mathrm{P}(C)\prod_i \mathrm{P}(Q_i \mid Q_{i-n+1}, \cdots, Q_{i-1}; C) \tag{5}$$

Here a class-specific n-gram model is used to model $\mathrm{P}(Q \mid C)$. The n-gram model parameters can be estimated with ML on a labeled training set. An n-gram model is often smoothed by interpolating with a lower order model. For the experiments in this paper, we used the interpolation of unigram and bigram models:

$$\mathrm{P}(Q \mid C) = \prod_i [(\delta \mathrm{P}(Q_i \mid C) + (1-\delta)\mathrm{P}(Q_i \mid Q_{i-1}, C)] \tag{6}$$

The n-gram classification model is also used for information retrieval when each document in a document collection is treated as a class c [10].

2.3. TF*IDF Weighted Vector Space Model

The TF*IDF weighted vector space model is widely used in information retrieval (IR). It represents a query (document) with a vector q (d). The relevance (or similarity) of the document to the query is measured as the cosine between the two vectors:

$$\cos(q,d) = \frac{q \cdot d}{\|q\|\|d\|} \tag{7}$$

For a document d, each element of its vector is a weight that represents the importance of a term (e.g., a word or a bigram) in the document. Intuitively, the importance should increase proportionally to the number of times a term appears in d and decreases when the term appears in many different documents. The *term frequency* $\mathrm{tf}_i(d)$ (TF) is the relative frequency of term i in d; the *inverse document frequency* (IDF) is the logarithm of the total number of documents divided by the number of documents containing i:

$$\mathrm{tf}_i(d) = \frac{n_i(d)}{\sum_k n_k(d)}, \quad \mathrm{idf}_i = \log \frac{|D|}{|\{d : i \in d\}|}$$

where $n_i(d)$ is the number of occurrences of term i in d, and D is the entire document collection. The weight for term i in the vector is the product of its TF and IDF scores. The vector for a query can be defined similarly.

TF*IDF was originally proposed as a heuristic weighting scheme for terms in a query/document. The heuristic works extraordinarily well and is difficult to beat. This leads to many theoretical justifications for the weighting scheme, as introduced in [7]. A justification that is closely related to the work in the paper can be found in [8]. It relates the TF*IDF weighting scheme to the MaxEnt model. It first generalizes

the MaxEnt by removing the restriction that the objective function is a probabilistic distribution and replacing it with a nonnegative function. It does so by introducing a KL-divergence generalized to nonnegative functions. It shows that the TF*IDF weight is the optimal weight of a feature in such a generalized model (simplified to contain a single term feature to enable a closed-form solution) in an information retrieval setting where each document is treated as the query that retrieves itself.

Strictly speaking, a TF*IDF model is just a matrix that measures the similarity between two entities (e.g., between a query and a document) with its originally usage in IR. We will show that it can be formalized as a classification model in the following session.

3. MAXENT MODEL PARAMETERIZATION

While the MaxEnt model has a convex objective function hence a global optimum regardless of the initial parameter setting, model initialization can still be an important issue due to the early stopping of training and the different settings of hyper-parameters for model regularization. This session describes how the parameters from an n-gram classification model or a TF*IDF model can be imported by a MaxEnt model for model initialization and hyper-parameter setting.

3.1. Linear Models

The N-gram classifiers, TF*IDF and MaxEnt models all have classification boundaries linear to the feature functions. This section studies how the decision functions of the n-gram classification and the TF*IDF model can be explicitly expressed as the linear combination of the classification features, with the focus on class prior, unigram and bigram features that are commonly used in text classification. The coefficients of these features can be imported by the MaxEnt model for initialization or hyper-parameter setting.

3.1.1. N-gram MaxEnt Initialization

Eq. (5) can be written with respect to each term t and term bigram *ht* in the query:

$$\begin{aligned}
&\log \mathrm{P}(c \mid q) \\
&= \log \mathrm{P}(c) + \sum_{ht} N(ht;q) \log\left(\delta \mathrm{P}(t \mid c) + (1-\delta)\mathrm{P}(t \mid h,c)\right) \\
&= \log \mathrm{P}(c) + \sum_{t} N(t;q) \log\left(\delta \mathrm{P}(t \mid c)\right) \\
&\quad + \sum_{ht} N(ht;q) \log\left(1 + \frac{(1-\delta)\mathrm{P}(t \mid h;c)}{\delta \mathrm{P}(t \mid c)}\right) \\
&= f_c(c,q) \log \mathrm{P}(c) + \sum_{t} f_{c,t}(c,q) \log\left(\delta \mathrm{P}(t \mid c)\right) \\
&\quad + \sum_{ht} f_{c,ht}(c,q) \log\left(1 + \frac{(1-\delta)\mathrm{P}(t \mid h;c)}{\delta \mathrm{P}(t \mid c)}\right)
\end{aligned} \tag{8}$$

In the last step of Eq. (8), $N(t;q)$ and $N(ht;q)$, i.e., the unigram and bigram counts in q, are written as the value of integer unigram and bigram feature functions $f_{c,t}$ and $f_{c,ht}$. f_c is the class prior feature:

$$f_c(C,Q) = \begin{cases} 1 & \text{if } C = c \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

According to Eq. (8), $\log \mathrm{P}(c)$ should be the weight for the class prior feature f_c; $\log\left(\delta \mathrm{P}(t \mid c)\right)$ is the weight for the unigram feature $f_{c,t}$; and $\log\left(1 + \frac{(1-\delta)\mathrm{P}(t \mid h;c)}{\delta \mathrm{P}(t \mid c)}\right)$ is the weight for the bigram feature $f_{c,ht}$.

*3.1.2. TF*IDF Vector Space MaxEnt Initialization*

In a TF*IDF model, the cosine score between a class c and a query q in Eq. (7) can be written with respect to each term t (unlike in the previous section, here t represents both unigrams and bigrams in the query. See [8] for the proper calculation of bigram IDF values):

$$\begin{aligned}
\cos(q,c) &= \frac{q \cdot c}{\|q\|\|c\|} = \sum_{t \in q} \frac{\mathrm{tf}_t(q) \times \mathrm{idf}_t \times \mathrm{tf}_t(c) \times \mathrm{idf}_t}{\|q\|\|c\|} \\
&= K \sum_{t \in q} \frac{f_{c,t}(c,q) \times \mathrm{tf}_t(c) \times \mathrm{idf}_t^2}{\|c\|}
\end{aligned} \tag{10}$$

Here since the norm of the query does not affect the classification boundary, it gets absorbed by the constant factor K. The relative term frequency $\mathrm{tf}_t(q)$ is replaced by the integer feature value (the number of occurrences of a term) $f_{c,t}(c,q)$ because they differ by a constant factor – the number of occurrences of all different terms. Because K does not change the decision boundary, the weight in this linear classification model for the feature $f_{c,t}(c,q)$ can be

$$\lambda_{c,t} = \mathrm{tf}_t(c)\,\mathrm{idf}_t^2 / \|c\| \tag{11}$$

Eq. (11) can be viewed as a parameter sharing mechanism. While there are $|C| \times |T|$ parameters in a linear classification model, they all depend on $\mathrm{tf}_t(c)$, idf_t, and $\|c\|$. There are only $|T|$ and $|C|$ parameters for the IDFs and the class norms. And the term frequency parameters depend only on the rank of a term in a class instead of its identity. Therefore all the terms having the same rank in a class (document) have their parameters tied. Given the fact that the number of different term ranks is much smaller than the number of different terms, this may greatly reduce the number of free parameters and improve its robustness.

3.2. MaxEnt Scaling

For a linear classification model, scaling of its parameters by a constant factor will not change the decision boundary.

However, the scaling of model parameters will change the value of the MaxEnt objective function. Although the theoretical background has been laid in [8] that motivates this work, it has made many pre-assumptions that does not apply in practical problems. The initial parameterization is hardly in the optimal scale for the MaxEnt objective function. Therefore we have to first scale the initialization to optimize the MaxEnt objective function after it has been imported from another linear classifier. Formally, we need to find the scaling factor k that maximize

$$\mathrm{P}(C \mid Q) = \frac{1}{Z_\lambda(Q)} \exp\left(\sum_{f_i \in \mathcal{F}} k\lambda_i f_i(C,Q)\right) \quad (12)$$

with the λ parameters fixed at their imported values. This can be done with a gradient based optimization, where

$$\frac{\partial \log \mathrm{P}(C \mid Q)}{\partial k} = \mathsf{E}_{\hat{\mathrm{P}}(Q,C)} \sum_{f_i} \lambda_i f_i(C,Q) - \mathsf{E}_{\hat{\mathrm{P}}(Q)\mathrm{P}(C|Q)} \sum_{f_i} \lambda_i f_i(C,Q) \quad (13)$$

3.3. MaxEnt Regularization Hyper-Parameter Setting

Instead of using zero means for the Gaussian priors in Eq. (4), m_i can be initialized with another linear classifier's (oracle's) parameterization. In doing so, the regularization takes into account the importance of features determined by a simpler (with fewer free parameters) model instead of treating them equally.

4. EXPERIMENTAL RESULTS

4.1. Experimental Settings

We conducted experiments with two different data sets, the Air Travel Information System data (ATIS) [11] in the public domain and a Microsoft internal product review sentiment classification data set. ATIS was originally not a classification task. We followed the practice in [12] to use the data for call-routing experiments by assigning the main database table name in the manually created SQL query (available in the NIST ATIS data set) for an utterance as its classification destination. There are 14 classes in total. In ATIS2 and ATIS3 dataset, 4995 training sentences, 828 development sentences and 914 test sentences are available for our experiments. We used unigram and bigram terms in MaxEnt modeling, which yields 98182 distinct features. The sentiment classification is a binary classification task with two destination classes (positive versus negative.) 22488 training, 4817 development and 4821 test examples are available. Unigram and bigram terms are used, which yields 520,740 features.

To train the TF*IDF model, we concatenate all examples with the same destination class to form a "document", and a TF*IDF weighted vector is constructed to represent the class. Similarly, all the examples labeled with the same destination class are pooled together to train the class specific n-gram model for the n-gram classifier.

For the MaxEnt parameterization with a TF*IDF model, we compare the classification accuracies in five different settings. The baseline uses the flat initialization where all the model parameters and the means for the regularization Gaussians are set to 0. The TF*IDF initialization sets the initial model parameters according to Eq. (11) and the regularization Gaussian means to 0. The scaled TF*IDF initialization sets the MaxEnt parameters according to Eq. (11) and then scales the parameters by a factor of k found by optimizing the objective function in Eq. (12). The "TF*IDF mean" setting sets not only the initial parameters but also the regularization Gaussian means according to Eq. (11). The "scaled TF*IDF mean" setting sets the parameters and the Gaussian means to the scaled values. Similarly, we compared the MaxEnt parameterization with an n-gram classifier in five settings – flat initialization, non-scaled and scaled initialization according to Eq. (8), non-scaled and scaled initialization/regularization mean setting according to Eq. (8).

For each setting, experiments were conducted with five different variances σ^2 for the regularization prior – ∞ (no regularization), 10, 20, 30 and 40. The MaxEnt models are trained with early stopping based on the posterior class probability on the development sets after the initialization.

4.1. Results

Table 1 compares the accuracies of the MaxEnt models with different TF*IDF parameterization settings on the ATIS test data. The accuracies are either lowered or not significantly improved when non-scaled TF*IDF parameters was used to initialize/regularize the MaxEnt model. With proper scaling, both TF*IDF initialization and regularization gets better accuracies – three out of five are statistically significant with the TF*IDF initialization, and four out of five with the TF*IDF regularization. The best TF*IDF initialized model (across different prior variances) improves the best baseline result significantly too. Table 2 shows the results with n-gram parameterizations for the MaxEnt model. Again, it shows that initialization/regularization without proper parameter scaling does not improve the accuracies, while fewer statistically significant improvements are observed when the parameters are properly scaled, and no significant improvement has been observed from the best results over all prior variances.

We also tried to train the models without early stopping according to a cross-validation set. The training "converges" when the difference of the average log conditional probability of the training data between two adjacent steps is smaller than 10^{-7}. It took 40% more time for the flat initialization to "converge." However, the flat and TD*IDF

initializations did not "converge" to the same point. This can be attributed to a very flat region in the parameter space. The test set accuracy is 94.30% for the flat initialization and 95.07% for the TF*IDF initialization.

σ^2	Flat init.	TF*IDF init.	TF*IDF init. (scaled)	TF*IDF mean	TF*IDF mean (scaled)
∞	94.52%	94.41%	95.18%	94.41%	95.18%
40	94.30%	94.30%	95.29%*†	94.30%	95.40%*†
30	94.19%	94.30%	95.29%*†	94.19%	95.40%*†
20	94.19%	93.87%	94.96%	94.52%	95.29%*
10	93.98%	94.09%	94.96%*	94.41%	95.18%*

Table 1. Bigram Maxent accuracies on the ATIS test data with the TF*IDF parameterizations: the "σ^2" column indicates the variance of the Gaussian for regularization. The asterisk indicates that the improvement over the baseline in the same row is statistically significant according to a sign test. † indicates that the best results across all σ^2 has significantly improved the best baseline result (94.52%) across all σ^2.

σ^2	Flat init.	N-gram init.	N-gram init. (scaled)	N-gram mean	N-gram mean (scaled)
∞	94.52%	94.63%	94.63%	94.63%	94.63%
40	94.30%	94.41%	94.52%	94.41%	94.52%
30	94.19%	94.74%	94.52%	94.41%	94.63%
20	94.19%	94.41%	93.98%	94.30%	94.96%*
10	93.98%	94.74%*	94.85%*	94.19%	94.96%*

Table 2. Bigram Maxent accuracies on the ATIS test data with the n-gram parameterizations. The asterisk indicates that the improvement over the baseline is statistically significant. No significant improvement has been observed when the best results across different rows are compared with the best baseline result.

σ^2	Flat init.	TF*IDF init.	TF*IDF init. (scaled)	TF*IDF mean	TF*IDF mean (scaled)
∞	76.80%	76.97%*	77.10%*	76.97%*	77.10%*
40	76.85%	76.97%	77.10%*	76.97%	77.07%*
30	76.87%	76.97%	77.10%*	76.97%	77.07%
20	76.93%	76.97%	77.12%	76.99%	77.12%
10	76.83%	77.01%*	77.14%*	77.03%*	77.12%*

Table 3. Bigram Maxent accuracies on the sentiment test data with the TF*IDF parameterizations. The asterisk in a cell indicates that the improvement over the baseline is statistically significant according to a sign test. No significant improvement has been observed when the best results across different rows are compared with the best baseline result across different rows.

Similar pattern is observed with the sentiment classification, where the TF*IDF initialization/regularization results in statistically significant improvement in accuracies in most cases when parameters are properly scaled (Table 3). However, the best TF*IDF parameterized model across all σ^2 does not exhibit significant improvements over the best baseline model. Without early stopping according to the cross validation data, the flat and TF*IDF initialized model with no prior converges to the same point, with the test set accuracy at 76.80%. No signification improvements have been observed with the n-gram parameterizations at all (Table 4). In contrast to the ATIS experiments, the effect of regularization is less significant due to the fact that the task has fewer destination classes but much more training data. Hence the data sparsity is less an issue than the ATIS classification task.

σ^2	Flat init.	N-gram init.	N-gram init. (scaled)	N-gram mean	N-gram mean (scaled)
∞	76.80%	76.80%	76.80%	76.80%	76.80%
40	76.85%	76.85%	76.85%	76.85%	76.85%
30	76.87%	76.87%	76.87%	76.87%	76.87%
20	76.93%	76.93%	76.93%	76.93%	76.93%
10	76.83%	76.83%	76.83%	76.83%	76.83%

Table 4. Bigram Maxent accuracies on the sentiment test data with the n-gram parameterizations. No statistically significant improvement has been observed.

σ^2	Flat init.	TF*IDF init.	TF*IDF init. (scaled)	TF*IDF mean	TF*IDF mean (scaled)
∞	94	91	494	83	509
40	113	104	244	69	428
30	231	88	192	60	389
20	111	86	174	113	187
10	101	86	116	146	194

Table 5. Training time (in seconds) of different TF*IDF MaxEnt parameterization settings with the ATIS data.

σ^2	Flat init.	N-gram init.	N-gram init. (scaled)	N-gram mean	N-gram mean (scaled)
∞	94	88	282	104	209
40	113	65	233	82	339
30	231	92	273	64	222
20	111	99	259	64	184
10	101	97	257	118	269

Table 6. Training time (in seconds) of different n-gram MaxEnt parameterization settings with the ATIS data.

σ^2	Flat init.	TF*IDF init.	TF*IDF init. (scaled)	TF*IDF mean	TF*IDF mean (scaled)
∞	2746	2908	2161	2069	2322
40	2575	3106	3438	2602	3433
30	2233	2448	3253	2070	2425
20	2224	2853	2404	2533	3006
10	2254	2477	2914	2469	3623

Table 7. Training time (in seconds) of different TF*IDF MaxEnt parameterization settings with the sentiment classification data.

σ^2	Flat init.	N-gram init.	N-gram init. (scaled)	N-gram mean	N-gram mean (scaled)
∞	2746	2039	2519	2563	2448
40	2575	2803	2233	1870	2918
30	2233	3019	2243	2928	2541
20	2224	2918	2467	3130	2466
10	2254	3014	2963	2404	2528

Table 8. Training time (in seconds) of different n-gram MaxEnt parameterization settings with the sentiment classification data.

Table 5~Table 8 compare the training time of different tasks with different parameterization settings. For the ATIS task, the scaling of the imported parameters greatly increases the training time, while the direct import without scaling does not have an obvious impact on the training speed. The impact of model scaling on training speed is less obvious with the sentiment classification data, where the much bigger training set and feature space make the MaxEnt training after initialization take much longer time, so the fraction of time spent on initial parameter scaling is much smaller in the entire training process.

5. DISCUSSIONS AND CONCLUSIONS

The TF*IDF weighted vector space model is very robust in comparing the similarity between a query and a document. In terms of text classification, each document forms a class of its own and the model assigns a class to a query according to their similarity. In this case, there is only one example for each class. The robustness (may be partly attributed to the weight sharing mechanism described in subsection 3.1.2) and the discriminative power of the TF*IDF model makes it difficult to beat. On the other hand, the MaxEnt model works much better in text classification when there are much more examples available for each class. When data are sparse but more than one per class, the TF*IDF parameterization improves the robustness of the MaxEnt model. This explains the more significant improvements on the ATIS data than the sentiment data, since the latter has much more examples per class.

We have shown that the initialization and hyper-parameter setting can have a significant impact on the test set accuracy of a MaxEnt model. While the TF*IDF initialization and hyper-parameter setting for MaxEnt models have improved the classification accuracy on test data significantly, less improvements have been observed from the n-gram MaxEnt initialization/hyper-parameter setting. We believe this may be attributed to the robustness and the discriminative nature of the TF*IDF vector space model and its parameter tying mechanism that reduces the number of free parameters. We have also discovered that the proper scaling of the initialization parameters is crucial in achieving the gains in classification accuracy.

6. REFERENCES

[1] R. Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling," *Computer Speech and Language*, vol. 10, pp. 187-228, 1996.

[2] Y.-Y. Wang, J. Lee, and A. Acero, "Speech Utterance Classification Model Training Without Manual Transcriptions," in the proceedings of the *International Conference on Acoustics, Speech, and Signal Processing*: IEEE, 2006.

[3] C. White, J. Droppo, A. Acero, and J. Odell, "Maximum Entropy Confidence Estimation for Speech Recognition," in the proceedings of the *International Conference on Acoustics, Speech, and Signal Processing*: IEEE, 2007.

[4] Y.-Y. Wang, D. Yu, Y.-C. Ju, G. Zweig, and A. Acero, "Confidence Measures for Voice Search Applications," in the proceedings of *Eurospeech*. Antwerp, Belgium: ISCA, 2007.

[5] J. H. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Ann. Math. Stat.,*, vol. 43, pp. 1470-1480, 1972.

[6] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*: Springer-Verlag, 1997.

[7] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF" *Journal of Documentation* vol. 60, pp. 503 - 520 2004.

[8] K. Papineni, "Why inverse document frequency?," in the proceedings of *North American Chapter Of The Association For Computational Linguistics*, 2001.

[9] A. L. Berger, S. A. D. Pietra, and V. J. D. Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, vol. 22, pp. 39-72, 1996.

[10] J. M. Ponte, W. B. Croft, and "A Language Modeling Approach to Information Retrieval," in the proceedings of *the 21st annual international ACM SIGIR conference*. Melbourne, Australia 1998, pp. 275 - 281

[11] P. Price, "Evaluation of Spoken Language System: the ATIS domain," in the proceedings of the *DARPA Speech and Natural Language Workshop*, Hidden Valley, PA, 1990.

[12] C. Chelba, M. Mahajan, and A. Acero, "Speech Utterance Classification," in the proceedings of the *International Conference on Acoustics, Speech, and Signal Processing*: IEEE, 2003.