



Minimum Rank Error Training for Language Modeling

Meng-Sung Wu and Jen-Tzung Chien

Department of Computer Science and Information Engineering
 National Cheng Kung University, Tainan, Taiwan 70101, ROC
 {mswu, chien}@chien.csie.ncku.edu.tw

Abstract

Discriminative training techniques have been successfully developed for many pattern recognition applications. In speech recognition, discriminative training aims to minimize the metric of word error rate. However, in an information retrieval system, the best performance should be achieved by maximizing the *average precision*. In this paper, we construct the discriminative *n*-gram language model for information retrieval following the metric of minimum rank error (MRE) rather than the conventional metric of minimum classification error. In the optimization procedure, we maximize the average precision and estimate the language model towards attaining the smallest ranking loss. In the experiments on ad-hoc retrieval using TREC collections, the proposed MRE language model performs better than the maximum likelihood and the minimum classification error language models.

Index Terms: discriminative training, minimum rank error, maximum average precision, language model, information retrieval

1. Introduction

Statistical language models have been successfully applied in many useful systems, including speech recognition, machine translation and information retrieval. However, speech recognition system is always evaluated by the word error rate of test data rather than the likelihood score of training data. For this concern, the discriminative learning methods have been extensively studied in the past decades and applied in many other systems [1][5][11]. The maximum mutual information (MMI) and minimum classification error (MCE) [1][2][8][9] criteria have been successfully applied to speech recognition and information retrieval. MMI method attempted to maximize the mutual information between training data and their correct word transcription. Using MCE procedure, the misclassification rate was approximated by a differentiable objective function and optimized by the generalized probabilistic descent (GPD) algorithm [9]. Kuo et al. [10] addressed that the discriminative training was able to improve language modeling for the application of speech recognition. Previous models [1][11] viewed the retrieval problem as a binary classification problem where the relevant query-document pairs were considered as positive data and the irrelevant query-document pairs were used as negative data. Also, a popular discriminative model was developed for information retrieval using the maximum entropy model [11]. Chen et al. [1] presented the MCE model training process via improving the discrimination among the hidden Markov models of different documents. However, in many situations, the classification error rate is not a suitable metric for measuring the rank of input document [4][6]. We are seeking the discriminative language model for information retrieval.

The performance of an information retrieval system is measured in terms of precision and recall. A popular measure that takes into account both recall and precision is the average precision [13]. In [5], the information retrieval model was trained by maximizing the average precision on training data. It was called the *MaxAP* training. Unfortunately, in this method, there was no analytical approach available for optimal retrieval performance through the user specified objective function, e.g. the average precision measured by the language model based retrieval system. Therefore, a discriminative training procedure is proposed here and aimed at attaining the highest average precision using the newly trained language model. In this paper, a new learning algorithm is explored to adjust language model parameters towards minimizing the rank error of training documents. We build a continuous objective function of average precision so that the optimization procedure is feasible to estimate the discriminative language model. This MRE procedure is different from MCE procedure and shown to be effective to achieve desirable retrieval performance. The average precision loss function is determined to represent the rank errors due to relevant and irrelevant documents.

2. Related Works

2.1. Language Model for Information Retrieval

Language model approach was first introduced for information retrieval by Ponte and Croft [12]. This approach is to build language model M_d for each document D and use it to rank D according to the probability of generating a query Q from the document model

$$P(Q|D) = P(w_1, w_2, \dots, w_K | M_d) = \prod_{w \in Q, k=1}^K P(w_k | M_d). \quad (1)$$

This unigram model can be easily extended to the *n*-gram model. Nevertheless, there is no reference of class variable C that denotes the relevance or the irrelevance. Given a query and a set of documents, the retrieval system ranks the documents based on the maximum *a posteriori* decision rule and finds the optimal document D for a given query

$$\hat{D} = \arg \max_D P(D|Q) = \arg \max_D P(Q|D)P(D), \quad (2)$$

where $P(Q|D)$ is the likelihood of a query given a document model, $P(D)$ is the prior probability that document D is relevant.

2.2. MCE Discriminative Training

In MCE training of acoustic models for speech recognition, a loss function [8][9] was calculated to approximate the classification error of speech recognizer. A misclassification measure was defined by

$$d_r(x) = -\log P(x|C_r) + \log \left(\frac{1}{C-1} \sum_{c_i, c_r \neq c_i} \exp(\log P(x|C_i)) \right)^\eta, \quad (3)$$

where η was a positive number, C was the number of classes, $P(x|C_r)$ and $P(x|C_i)$ were the likelihood functions of target (relevant) and competing (irrelevant) classes. The sigmoid function was used to define the class loss function [8][9] as

$$l(d_r(x)) = \frac{1}{1 + \exp(-\alpha d_r(x) + \beta)}, \quad (4)$$

where α and β were control parameters for the slope and the offset of the function, respectively. However, such a classification framework should be modified and followed by the metric used for information retrieval. Also, this misclassification error could not reflect the metric of average precision as a measure for information retrieval. Minimizing the classification error does not guarantee the high ranking of test documents. To deal with this issue, we develop a new learning algorithm of language model by optimizing the retrieval performance with the highest average precision.

2.3. Average Precision versus Classification Accuracy

Average precision is the average of precision computed after truncating the list after each of the relevant documents in turn. Let us consider two systems that produce probability estimates for a set of 10 documents. We assume that both systems retrieve five as relevant documents and the other five as irrelevant documents. We rank the test documents according to the probability of + (relevant), as shown in Table 1. If we calculate average precision, we obtain 1 $((1/1+2/2+3/3+4/4)/4)$ and 0.6792 $((1/2+2/3+3/4+4/5)/4)$ for retrieval system 1 and system 2, respectively. However, we see that two systems for binary classification have the same classification accuracy 80% and thus they are equivalent in terms of classification accuracy. Clearly, we know that system 1 is better than system 2 since system 1 shows a better overall ranking performance.

Table 1: Two systems have the same classification accuracy but different average precision [6]

Docs	d ₁	d ₂	d ₃	d ₄	d ₅	d ₆	d ₇	d ₈	d ₉	d ₁₀
sys_1	-	-	-	-	+	-	+	+	+	+
rank					6		4	3	2	1
sys_2	+	-	-	-	-	+	+	+	+	-
rank	10					4	3	2	1	

3. Minimum Rank Error Training

Average precision is the most frequently used in information retrieval. A perfect system can attain an average precision (AP) of 1. Therefore, we use 1-AP to represent the *ranking loss*. The optimization of the ranking loss is comparable to the optimization of the average precision. It is clear that minimum rank error is comparable to minimizing the number of irrelevant documents scored higher than the relevant documents. This consideration motivates us to use the rank error measure instead of a misclassification measure in information retrieval system.

3.1. MRE Discriminative Training

The relevant and irrelevant documents given a query can be evaluated by the probability of the query matching with the document language model. The training data of a retrieval system include a set of queries, a set of documents and the relevance judgments that manually label the pairs of queries and documents as relevance and irrelevance. A desirable retrieval model should be able to rank all relevant documents higher than irrelevant ones for a given query. The performance of system will not be satisfied if we cannot distinguish the difference between correct and incorrect answers. Typically, small rank error rate implies good performance. We attempt toward finding optimal language model parameters with minimum rank error (MRE). Therefore, we propose a new language model where local optimum of MRE criterion is achieved. In the first stage, a minimum rank error is optimized and closely related to the average precision performance. In the second stage, a learning rule is performed to optimize the criterion of average precision. Model parameters are adjusted according these ranked documents. The diagram for learning the language model parameters for information retrieval is displayed in Figure 1.

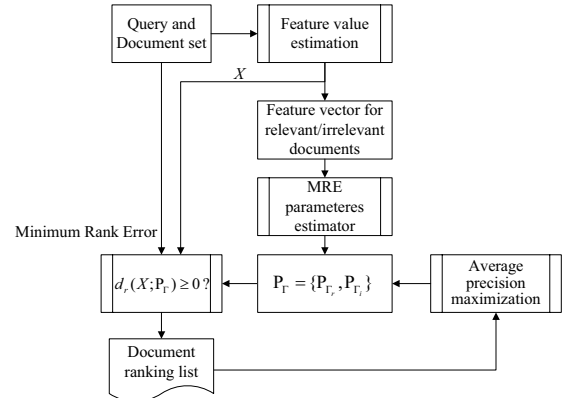


Figure 1: Discriminative information retrieval via average precision maximization

3.2. Maximum Average Precision & Minimum Rank Error

In a document retrieval system, X is the set of documents which answer the given queries Q . For each query, the training documents are comprised of a relevant set X_r and an irrelevant set X_i . The matching score of a document d with respect to a given query q can be defined by a discriminant function $f(x)$. Let $x_r \succ x_i$ denote that x_r is ranked higher than x_i . The objective function is defined by considering the classification rule

$$x_r \succ x_i \Leftrightarrow f(x_r) > f(x_i), \forall x_r \in S_R, x_i \in S_I. \quad (5)$$

The loss functions should be related to the number of rank errors that decision function f makes on the training set. The rank error rate [3][14] is the number of a lower scoring irrelevant document that is incorrectly ranked above the relevant document. This objective is determined by the probability of misordering a document pair (x_r, x_i) , i.e. occurring $f(x_i) > f(x_r)$. The rank errors are accumulated by

$$\begin{aligned}
Error(X) &= \sum_{x_i, x_r, i \neq r} P[f(x_i) > f(x_r)] \\
&= \sum_{q_t \in Q} \sum_{x_i, x_r, i \neq r} I[-f(q_t, x_r; P_\Gamma) + f(q_t, x_i; P_\Gamma)] \quad (6) \\
&= \sum_{q_t \in Q} \sum_{x_i, x_r, i \neq r} I[-P(q_t | x_r; P_\Gamma) + P(q_t | x_i; P_\Gamma)]
\end{aligned}$$

$$I[y] = \begin{cases} 1, & \text{for } y > 0 \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

where $I[y]$ is an indicator function and P_Γ is language model. In (6), the discriminant function is written by $f(x) = f(q_t, x; P_\Gamma) \equiv P(q_t | x; P_\Gamma)$. We can define the total distance measure by

$$d_r(X; P_\Gamma) = -\log P(Q | X_r; P_\Gamma) + \max_{d_i} \log P(Q | X_i; P_\Gamma) \quad (8)$$

which is the difference between the log likelihood scores of the target document and the most competing document. If this distance is negative, no rank error is induced. It is easy to see that minimizing (6) is equivalent to maximizing the average rank of the relevant documents. Let n_{xy} denote the number of class- x input tokens which are classified as class- y . Here, x, y have the labels r and i corresponding to the relevant state and the irrelevant state, respectively. Figure 2 illustrates the metric of average precision.

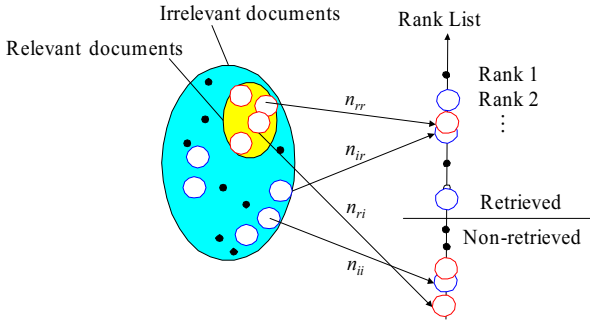


Figure 2: Illustration of the metric of average precision

Let N represent the number of relevant documents is retrieved. The precision p is given by

$$p = \frac{n_{rr}}{n_{rr} + n_{ir}} \quad (9)$$

and the average precision is defined by

$$AP = \frac{1}{N} \sum_{n=1}^N \frac{n_{rr}}{n_{rr} + n_{ir}} \quad (10)$$

where n_{rr} and n_{ir} are referred as the numbers of relevant and irrelevant documents retrieved as the relevance class, respectively. Maximizing the average precision is tightly related to minimizing the following ranking loss

$$L_{AP}(X; P_\Gamma) = 1 - AP = \frac{1}{N} \sum_{n=1}^N \left(1 - \frac{n_{rr}}{n_{rr} + n_{ir}} \right) = \frac{1}{N} \sum_{n=1}^N \left(\frac{n_{ir}}{n_{rr} + n_{ir}} \right) \quad (11)$$

In (11), the error count n_{ir} is the rank error determined by the misclassification measure $d_r(X; P_\Gamma)$. Minimizing the ranking loss is comparable of achieving maximum average precision. Similar to MCE algorithm, we express the ranking

loss function L_{AP} as a differentiable objective. The error count n_{ir} is approximated by the differentiable loss function l_{ir} defined as

$$n_{ir} \approx l_{ir} \equiv l(d_r(X, P_\Gamma)) \quad (12)$$

where $l(\cdot)$ is a sigmoid function in (4). The differentiation of the ranking loss function turns out to be

$$\frac{\partial L_{AP}(X; P_\Gamma)}{\partial P_\Gamma} = \frac{\partial L_{AP}}{\partial l_{ir}} \cdot \frac{\partial l_{ir}}{\partial d_r} \cdot \frac{\partial d_r}{\partial P_\Gamma} \quad (13)$$

where

$$\frac{\partial L_{AP}}{\partial l_{ir}} = \frac{1}{N} \sum_{n=1}^N \frac{n_{rr}}{(n_{rr} + n_{ir})^2} \quad (14)$$

and

$$\frac{\partial l_{ir}}{\partial d_r} = l(d_r(\cdot)) \cdot (1 - l(d_r(\cdot))) \quad (15)$$

Without loss of generality, we use a bigram language model as an example. Let $P_{m,n} = \log P(w_n | w_m, x)$ denote the bigram log probability for two-word event $w_{m,n} = (w_m, w_n)$. We have

$$\begin{aligned}
\frac{\partial d_r(X; P_\Gamma)}{\partial P_{m,n}} &= \sum_{q_t \in Q} \left[-n(x_r, w_{m,n}) P(q_t | x_r; P_\Gamma) \right. \\
&\quad \left. + \max_{d_i} n(x_i, w_{m,n}) P(q_t | x_i; P_\Gamma) \right] \quad (16)
\end{aligned}$$

where $n(x, w_{m,n})$ denotes the number of times the bigram $P_{m,n}$ appears in document x . Our language model P_Γ for the relevance class is to be adjusted to improve the metric of average precision. Using the steepest descent algorithm, the parameters of language model P_Γ are adjusted iteratively by

$$P_\Gamma(t+1) = P_\Gamma(t) - \varepsilon \cdot \frac{\partial L_{AP}(X; P_\Gamma(t))}{\partial P_\Gamma(t)} \quad (17)$$

where ε is a learning rate and t is an iteration index.

4. Experiments

In the experiments, we evaluate the performance of proposed MRE training of language model using two TREC collections. One is the Associated Press Newswire (1988) (AP88) dataset consists of 79,919 documents and the other dataset is the Wall Street Journal (1987) (WSJ87) consists of 46,448 documents.

4.1. Evaluation of Model Perplexity

When evaluating model perplexity, we used Wall Street Journal dataset as training data for language model estimation. The Associated Press Newswire dataset is used as the test data. For comparison, we carried out the baseline language model using maximum likelihood (ML) estimation. Discriminative language model via MRE training was realized. During MRE training procedure, we adopted the parameters $\eta = 1$, $\alpha = 1$, $\beta = 0$ and $\varepsilon = 0.5$. Table 2 reports the perplexity for unigram and bigram language models based on ML and MRE training. The baseline unigram and bigram obtain the perplexity of 1781.03 and 443.55, respectively. The perplexity is reduced to 1775.65 of unigram and 440.38 of bigram, which are trained using MRE algorithm. The perplexities of two methods are comparable because MRE discriminative training does not aim to improve the ability of language model to predict unseen texts. MRE

aims to maximize average precision rather than improve the performance of data fitting.

Table 2: Comparison of perplexity using ML and MRE language models

	ML	MRE
Unigram	1781.03	1775.65
Bigram	443.55	440.38

4.2. Experimental Results on Information Retrieval

In the next set of experiments, we evaluate the performance of discriminative training on the TREC ad-hoc information retrieval task. There were two query sets and the corresponding relevant documents in this collection. We used TREC topics 51-100 as training queries and the TREC topics 101-150 as test queries. Queries were sampled from the ‘title’ and ‘description’ fields of the topics. The average length of these queries contained about 8-12 words. In both training and test sets, we used ML language model as the baseline system. In each query test, we retrieved 1000 documents and calculated the average precision over all topics. To test the significance of improvement between two methods, Wilcoxon test [7] was employed in the evaluation. Table 3 compares the average precision on two datasets using ML and MRE training. For both datasets, MRE outperforms ML. Average precision is improved by 10.9% and 13.1% on two data. Table 4 highlights the evaluation of precision in document level using AP88 dataset. MCE training is also compared. R-precision is a measure of precision after R documents are retrieved where R is the number of relevant documents in the collection for a query. We find that a significant increase in averaged precision is obtained when comparing MRE over ML. The Wilcoxon test specifies the significance of improvement in precision is at the 5% threshold level. In some cases, MRE is not significantly better than MCE.

Table 3: Comparison of average precision using two TREC datasets

Collection	ML	MRE	Improvement	Wilcoxon
WSJ87	0.1012	0.1122	10.9%	0.0163*
AP88	0.1692	0.1913	13.1%	0*

Table 4: Comparison of precision in document level using AP88. All statistically significant performance improvements (Wilcoxon<0.05) are marked by stars.

Documents Retrieved	ML	MCE	MRE	Wilcoxon (MRE→ML)	Wilcoxon (MRE→MCE)
5 docs	0.416	0.464	0.472	0.0275*	0.1163
10 docs	0.406	0.426	0.456	0.0208*	0.0449*
15 docs	0.387	0.409	0.419	0.0251*	0.0447*
20 docs	0.373	0.397	0.405	0.0239*	0.0656
30 docs	0.339	0.361	0.371	0.0030*	0.0330*
100 docs	0.237	0.254	0.255	0*	0.0561
200 docs	0.165	0.173	0.175	0*	0.0622
500 docs	0.083	0.086	0.087	0*	0.0625
1000 docs	0.045	0.046	0.046	0*	0.0413*
R-Precision	0.232	0.247	0.273	0*	0.0096*

5. Conclusions

This paper presented a new discriminative training approach to language modeling based on the minimum rank error criterion. This criterion was a critical objective for document

retrieval. In previous studies, most of classification systems were designed by minimizing the classification errors, therefore they should not suitable when they were applied in ranking the retrieved documents. In this study, we estimated language model parameters towards optimizing the average precision objective function. Consequently, we constructed a novel discriminative retrieval model. In this new training procedure, we accumulated the rank errors, minimized the resulting average precision and came up with a learning algorithm of language model for information retrieval. The experiments on evaluating model perplexity and average precision illustrated the superiority of MRE training to ML training. In the future, we will extend this method for spoken document retrieval. Also, we are developing the alternative ranking criterion using the area under of the ROC curve (AUC) [4][6] for language modeling.

6. References

- [1] B. Chen, H.-M. Wang, and L.-S. Lee, “A discriminative HMM/N-Gram-Based retrieval approach for Mandarin spoken documents”, *ACM Trans. Asian Language Information Processing*, vol. 3, no. 2, pp. 128-145, 2004.
- [2] W. Chou, “Discriminant-function-based minimum recognition error rate pattern-recognition approach to speech recognition”, *Proceedings of the IEEE*, vol. 88, pp. 1201-1223, 2000.
- [3] M. Collins, “Discriminative reranking for natural language parsing”, in *Proc. 17th International Conference on Machine Learning*, pp. 175-182, 2000.
- [4] C. Cortes and M. Mohri, “AUC optimization vs. error rate minimization”, in *Advances in Neural Information Processing Systems*, vol. 15, 2003.
- [5] J. Gao, H. Qi, X. Xia, J.-Y. Nie, “Linear discriminant model for information retrieval”, in *Proc. ACM SIGIR*, pp.290-297, 2005.
- [6] J. Huang and C. X. Ling, “Using AUC and Accuracy in Evaluating Learning Algorithms”, *IEEE Trans. Knowledge and Data Engineering*, vol. 17, no. 3, pp.299-310, 2005.
- [7] D. Hull, “Using statistical testing in the evaluation of retrieval experiments”, in *Proc ACM SIGIR*, pp. 329-338, 1993.
- [8] B. H. Juang, W. Chou, and C.-H. Lee, “Minimum classification error rate methods for speech recognition”, *IEEE Trans. Speech and Audio Processing*, pp. 257-265, 1997.
- [9] B.-H. Juang and S. Katagiri, “Discriminative learning for minimum error classification”, *IEEE Trans. Signal Processing*, vol. 40, no. 12, pp. 3043-3054, 1992.
- [10] H.-K. J. Kuo, E. Fosler-Lussier, H. Jiang, and C.-H. Lee, “Discriminative training of language models for speech recognition”, in *Proc. ICASSP*, pp. 325-328, 2002.
- [11] R. Nallapati, “Discriminative models for information retrieval”, in *Proc. ACM SIGIR*, pp. 64-71, 2004.
- [12] J. M. Ponte and W. B. Croft, “A language modeling approach to information retrieval”, in *Proc. ACM SIGIR*, pp.275-281, 1998.
- [13] G. Salton, and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.
- [14] J.-N. Vittaut and P. Gallinari, “Machine learning ranking for structured information retrieval”, in *Proc. 28th European Conference on IR Research*, pp.338-349, 2006.