LATENT DIRICHLET LANGUAGE MODEL FOR SPEECH RECOGNITION

Jen-Tzung Chien and Chuang-Hua Chueh

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

ABSTRACT

Latent Dirichlet allocation (LDA) has been successfully presented for document modeling and classification. LDA calculates the document probability based on bag-of-words scheme without considering the sequence of words. This model discovers the topic structure at document level, which is different from the concern of word prediction in speech recognition. In this paper, we present a new latent Dirichlet language model (LDLM) for modeling of word sequence. A new Bayesian framework is introduced by merging the Dirichlet priors to characterize the uncertainty of latent topics of *n*-gram events. The robust topic-based language model is established accordingly. In the experiments, we implement LDLM for continuous speech recognition and obtain better performance than probabilistic latent semantic analysis (PLSA) based language method.

Index Terms—Natural languages, Bayes procedures, clustering methods, smoothing methods, speech recognition

1. INTRODUCTION

The statistical *n*-gram language models have been successfully developed for continuous speech recognition and many other applications. The *n*-gram model suffers from the insufficiencies of training data and long distance information, which limits the model generalization. The association pattern language model was presented to explore long distance associations of multiple words and merge them in *n*-gram model [3]. This insufficiency was also compensated by extracting the latent semantic information for topic-based language modeling. The latent semantic analysis (LSA) language model [1] was built accordingly. LSA performed matrix decomposition and found latent semantic information for different words and documents.

Hofmann [6] proposed the probabilistic LSA (PLSA), where the latent topic parameters were estimated by maximum likelihood method via the EM algorithm. The semantic regularities in *n*-gram events were exploited for speech recognition [4][8]. In general, PLSA characterized the collected documents individually by different parameters. Parameter size was increased significantly. The unseen documents could not be predicted. Blei *et al.* [2] presented the latent Dirichlet allocation (LDA) and improved the PLSA model by merging Dirichlet priors for topic mixtures. Seen and unseen documents were consistently generated by the LDA parameters, which were estimated by variation inference method. The model complexity was controlled. The new documents were generalized. LDA has been successfully applied in document classification [2] and information retrieval [11]. More recently, LDA was employed for adaptation of language model [5][9]. Nonetheless, PLSA and LDA were developed by exploiting topic information at document level. The extracted topic statistics are not directly representative for speech recognition, where the latent topics of *n*-gram events should be concerned.

In this work, we endeavor to build the *latent Dirichlet language model* (LDLM), where the topic structure of LDA model is merged in generation of language model probability. Different from LDA for modeling document probability, LDLM characterizes the *n*-gram regularities from different documents. A Bayesian approach is presented to calculate the model parameters under the assumption of latent topics of *n*-gram events being Dirichlet distributed. The topic characteristics are expressed in conditional probability of a word given its history, which is consistent to the function of word prediction for speech recognition. This LDLM is smoothed by incorporating the topic information and estimated by maximizing the marginal likelihood of training data. The uncertainty of topic variables is compensated for robust language modeling.

2. LATENT DIRICHLET ALLOCATION

LDA extends the PLSA model by treating the latent topic of each document as a random variable. The number of parameters is controlled even through the size of training documents is increased significantly. Different from PLSA, LDA model is capable of calculating likelihood function of unseen documents. Typically, LDA is a generative probabilistic model for documents in text corpus. The documents are represented by the random latent topics, which are characterized by the distributions over words. The graphical representation of LDA is shown in Figure 1. The LDA parameters consist of $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_C]$ denotes the Dirichlet parameters of C latent topic mixtures, and $\boldsymbol{\beta}$ is a matrix with multinomial entry $\beta_{i,w} = p(w \mid c_i)$. Using LDA, the probability of an N-word document $\mathbf{w} = [w_1, w_2, \cdots, w_N]$ is calculated by the following procedure.

First, a topic mixture vector $\boldsymbol{\theta}$ is drawn from the Dirichlet distribution with parameter $\boldsymbol{\alpha}$. The corresponding topic sequence $\mathbf{c}=[c_1,c_2,\cdots,c_N]$ is generated based on the multinomial distribution with parameter $\boldsymbol{\theta}$ in document level. Each word w_n is generated by the distribution $p(w_n \mid c_n,\boldsymbol{\beta})$. The joint probability of $\boldsymbol{\theta}$, topic assignment $\mathbf{c}$ and document $\mathbf{w}$ is given by

$$p(\boldsymbol{\theta},\mathbf{c},\mathbf{w} \mid \boldsymbol{\alpha},\boldsymbol{\beta}) = p(\boldsymbol{\theta} \mid \boldsymbol{\alpha})\prod_{n=1}^{N} p(c_n \mid \boldsymbol{\theta})p(w_n \mid c_n,\boldsymbol{\beta}) . \quad (1)$$

By integrating (1) over $\boldsymbol{\theta}$ and summing it over $\mathbf{c}$, we obtain the marginal probability of document $\mathbf{w}$ by

$$p(\mathbf{w} \mid \boldsymbol{\alpha},\boldsymbol{\beta}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{\alpha})\prod_{n=1}^{N}\sum_{c_n=1}^{C} p(c_n \mid \boldsymbol{\theta})p(w_n \mid c_n,\boldsymbol{\beta})d\boldsymbol{\theta} . \quad (2)$$

The LDA parameters $\{\boldsymbol{\alpha},\boldsymbol{\beta}\}$ are estimated by maximizing the marginal likelihood of training documents. The parameter estimation was solved by approximate inference algorithms including Laplace approximation, variational inference, and resampling method [7]. Using variational inference, the variational parameters were adopted for calculating the lower bound of marginal likelihood. The LDA parameters were estimated by maximizing the lower bound, or equivalently minimizing the Kullback-Leibler divergence between the variational distribution and the true posterior $p(\boldsymbol{\theta},\mathbf{c} \mid \mathbf{w},\boldsymbol{\alpha},\boldsymbol{\beta})$.

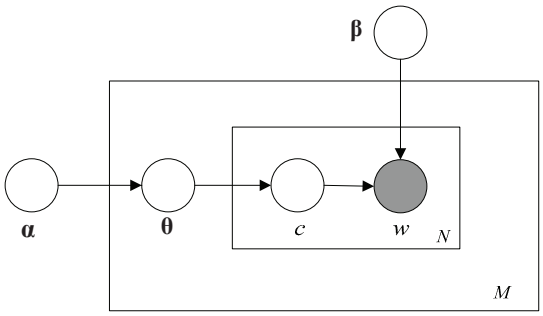

Figure 1 Graphical model of LDA

3. LATENT DIRICHLET LANGAUGE MODEL

LDA has been presented for document modeling and text categorization. In speech and language applications, LDA was implemented for language model adaptation [5]. LDA was used to calculate the document probability based on the bag-of-words scheme without taking the order of word occurrence into account. In addition, LDA was constructed by the topic structure at document level, which was inconsistent to the language model in speech recognition where the *n*-gram regularities were characterized. To develop LDA based topic language model, Wallach [10] presented the LDA bigram by considering the bigram events rather than unigram events in a document. Each document was seen as a bag of *bigrams*. The Gibbs sampling method was applied for parameter estimation. LDA bigrams were constructed at document level and shown to be effective for document modeling. Nevertheless, this method does not conform to the sentence generation in speech recognition system. Accordingly, we present a new latent Dirichlet language model (LDLM) for modeling of *n*-gram events for speech recognition. The topic structure is embedded in generation of *n*-gram probability. The uncertainty of topic information is compensated by integrating the likelihood function over the Dirichlet priors. LDLM parameters are estimated by maximizing the marginal likelihood.

3.1. Model construction

LDLM acts as the Bayesian extension of topic-based language model, where the prior information of topic variable is adopted. The graphical model of LDLM is illustrated in Figure 2. Here, H and V denote the number of histories and the size of vocabulary, respectively. Importantly, the *n*-gram event of a word and its history is represented by a new graphical model, where the topic mixture vector $\boldsymbol{\theta}$ is generated by the *history-dependent* Dirichlet parameter. By merging a parameter matrix $\mathbf{A}$, the Dirichlet model is generated by considering the word occurrence in each history. The history is represented by a $(n-1)V\times 1$ vector $\mathbf{h}$ consisting of $n-1$ block subvectors with the entries of the seen words are assigned by one and those of unseen words are zeros. The order of historical words is considered in $\mathbf{h}$. The observed *n*-gram event contains the current word w and its history $\mathbf{h}$. The topic mixture vector $\boldsymbol{\theta}$ is drawn by Dirichlet distribution

$$p(\boldsymbol{\theta} \mid \mathbf{h},\mathbf{A}) = \frac{\Gamma(\sum_{i=1}^{C}\mathbf{a}_i^T\mathbf{h})}{\prod_{i=1}^{C}\Gamma(\mathbf{a}_i^T\mathbf{h})}\theta_1^{\mathbf{a}_1^T\mathbf{h}-1}\cdots\theta_C^{\mathbf{a}_C^T\mathbf{h}-1}, \quad (3)$$

where $\mathbf{a}_i^T$ is the ith row of $\mathbf{A}$, $\Gamma(\cdot)$ is the gamma function and C is the number of topics. For each prediction word w, a topic c is chosen by a multinomial distribution with parameter $\boldsymbol{\theta}$. The probability of latent topic c and word w conditioned on history $\mathbf{h}$ is computed by

$$p(\boldsymbol{\theta},c,w \mid \mathbf{h},\mathbf{A},\boldsymbol{\beta}) = p(\boldsymbol{\theta} \mid \mathbf{h},\mathbf{A})p(c \mid \boldsymbol{\theta})p(w \mid c,\boldsymbol{\beta}) . \quad (4)$$

By integrating this probability over $\boldsymbol{\theta}$ and summing it over c, we obtain the conditional probability in language model

$$p(w \mid \mathbf{h},\mathbf{A},\boldsymbol{\beta}) = \int p(\boldsymbol{\theta} \mid \mathbf{h},\mathbf{A})\sum_{c} p(c \mid \boldsymbol{\theta})p(w \mid c,\boldsymbol{\beta})d\boldsymbol{\theta} . \quad (5)$$

LDLM is obtained accordingly. In comparison of LDLM in (5) and LDA in (2), LDLM calculates the word probability conditioned on historical context. LDA and LDLM generate the topic mixture vector $\boldsymbol{\theta}$ with the parameter $\boldsymbol{\alpha}$ and the combination of $\mathbf{A}$ and $\mathbf{h}$, respectively. LDLM is exploited to characterize the order of neighboring words and determine the language model for not only seen histories but also unseen histories. Through exploring the topic structure, those histories corresponding to the same topic assignment share the same word cluster distribution. This smoothed language model is assured for good generalization. LDLM is also referred as a Bayesian language model, where the

uncertainty of topic mixture vector $\boldsymbol{\theta}$ is incorporated for robust language modeling.

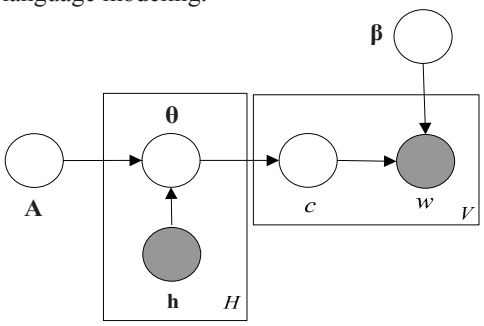

Figure 2 Graphical model of LDLM

3.2. Model estimation

The parameters $\{\mathbf{A},\boldsymbol{\beta}\}$ of LDLM can be estimated by maximizing the marginal distribution accumulated from all training data

$$\sum_{\mathbf{h},w} c(\mathbf{h},w)\log p(w\,|\,\mathbf{h},\mathbf{A},\boldsymbol{\beta}), \tag{6}$$

where $c(\mathbf{h},w)$ is the number of occurrence of n-gram event $(\mathbf{h},w)$ in training corpus. The LDLM $p(w\,|\,\mathbf{h},\mathbf{A},\boldsymbol{\beta})$ calculates the probability of a word w conditioned on its historical words $\mathbf{h}$. The integral over $\boldsymbol{\theta}$ in (5) is computed by

$$p(c_i\,|\,\mathbf{h},\mathbf{A}) = \int p(\boldsymbol{\theta}\,|\,\mathbf{h},\mathbf{A})\,p(c_i\,|\,\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{\mathbf{a}_i^T\mathbf{h}}{\sum_{j=1}^{C}\mathbf{a}_j^T\mathbf{h}}, \tag{7}$$

which is an expectation of Dirichlet distribution of a latent topic c_i. The variational inference is not needed at this circumstance. The probability of n-gram event using LDLM is expressed by

$$p(w\,|\,\mathbf{h},\mathbf{A},\boldsymbol{\beta}) = \sum_{i=1}^{C} p(w\,|\,c_i,\boldsymbol{\beta})\frac{\mathbf{a}_i^T\mathbf{h}}{\sum_{j=1}^{C}\mathbf{a}_j^T\mathbf{h}}. \tag{8}$$

Notably, the hidden variable c_i is used in marginal likelihood function $p(w\,|\,\mathbf{h},\mathbf{A},\boldsymbol{\beta})$. The problem of incomplete data $(\mathbf{h},w)$ happens in model estimation procedure. EM algorithm should be applied by considering complete data $(\mathbf{h},w,c)$. In E-step, the auxiliary function of new estimates $\{\mathbf{A}',\boldsymbol{\beta}'\}$ given current estimates $\{\mathbf{A},\boldsymbol{\beta}\}$ is calculated by taking the expectation of marginal likelihood function of (6) over the hidden variable c_i by

$$\begin{aligned} Q(\mathbf{A}',\boldsymbol{\beta}'\,|\,\mathbf{A},\boldsymbol{\beta}) &= \sum_{\mathbf{h},w} c(\mathbf{h},w)E_c[\log p(w,c\,|\,\mathbf{h},\mathbf{A}',\boldsymbol{\beta}')\,|\,\mathbf{A},\boldsymbol{\beta}] \\ &= \sum_{\mathbf{h},w} c(\mathbf{h},w)\sum_{i=1}^{C} p(c_i\,|\,\mathbf{h},w,\mathbf{A},\boldsymbol{\beta})\log p(w,c_i\,|\,\mathbf{h},\mathbf{A}',\boldsymbol{\beta}') \\ &= \sum_{\mathbf{h},w} c(\mathbf{h},w)\sum_{i=1}^{C} p(c_i\,|\,\mathbf{h},w,\mathbf{A},\boldsymbol{\beta})\log\left[p(w\,|\,c_i,\boldsymbol{\beta}')\frac{\mathbf{a}_i'^T\mathbf{h}}{\sum_{j=1}^{C}\mathbf{a}_j'^T\mathbf{h}}\right], \end{aligned} \tag{9}$$

where $p(c_i\,|\,\mathbf{h},w,\mathbf{A},\boldsymbol{\beta})$ is the posterior probability calculated by using the current estimate $\{\mathbf{A},\boldsymbol{\beta}\}$

$$p(c_i\,|\,\mathbf{h},w,\mathbf{A},\boldsymbol{\beta}) = \frac{p(w\,|\,c_i,\boldsymbol{\beta})\cdot\mathbf{a}_i^T\mathbf{h}}{\sum_{j=1}^{C}p(w\,|\,c_j,\boldsymbol{\beta})\cdot\mathbf{a}_j^T\mathbf{h}}. \tag{10}$$

In M-step, we maximize the auxiliary function with respect to the parameters $\beta'_{i,w} = p(w\,|\,c_i,\boldsymbol{\beta}')$ and $\mathbf{a}'_i$ and find the updated estimates $\{\mathbf{A}',\boldsymbol{\beta}'\}$. The Lagrange multipliers $\lambda = [\lambda_1,\cdots,\lambda_C]$ are applied in estimation of probability parameters $\boldsymbol{\beta}'$ where the extended objective function is established by

$$Q(\mathbf{A}',\boldsymbol{\beta}'\,|\,\mathbf{A},\boldsymbol{\beta}) + \sum_{i=1}^{C}\lambda_i\left[\sum_{w} p(w\,|\,c_i,\boldsymbol{\beta}') - 1\right]. \tag{11}$$

The new parameter $\beta'_{i,w}$ is updated by the closed-form formula

$$\beta'_{i,w} = \frac{\sum_{\mathbf{h}} c(\mathbf{h},w)\,p(c_i\,|\,\mathbf{h},w,\mathbf{A},\boldsymbol{\beta})}{\sum_{w'}\sum_{\mathbf{h}} c(\mathbf{h},w')\,p(c_i\,|\,\mathbf{h},w',\mathbf{A},\boldsymbol{\beta})}, \tag{12}$$

But, no closed-form solution exists for parameter $\mathbf{A}'$. The decent algorithm is employed to find $(n-1)V\times 1$ parameter vector $\mathbf{a}'_i$. To do so, the gradient function $\nabla_{\mathbf{a}'_i}Q(\mathbf{A}',\boldsymbol{\beta}'\,|\,\mathbf{A},\boldsymbol{\beta})$ is calculated by

$$\sum_{\mathbf{h},w} c(\mathbf{h},w)\,p(c_i\,|\,\mathbf{h},w,\mathbf{A},\boldsymbol{\beta})\left[\frac{1}{\mathbf{a}_i'^T\mathbf{h}} - \frac{1}{\sum_{j=1}^{C}\mathbf{a}_j'^T\mathbf{h}}\right]\mathbf{h}, \tag{13}$$

and the new parameter $\mathbf{a}_i'^{(t+1)}$ at $(t+1)$ iteration is updated by

$$\mathbf{a}_i'^{(t+1)} = \mathbf{a}_i'^{(t)} - \eta\nabla_{\mathbf{a}_i'^{(t)}}Q(\mathbf{A}',\boldsymbol{\beta}'\,|\,\mathbf{A},\boldsymbol{\beta}), \tag{14}$$

with learning rate η. LDLM parameters are estimated by several EM iterations.

3.3. Comparison of LDA, LDA bigram and LDLM

It is interesting to investigate the relations of LDA [2], LDA bigram [10] and LDLM. The LDA and LDA bigram endeavor to characterize the word distributions of documents, which contain bags of unigrams and bigrams, respectively. Their likelihood functions are calculated individually for training documents by using the shared Dirichlet parameter $\boldsymbol{\alpha}$. New documents are generated from the corresponding word collections. In contrast, LDLM focuses on exploiting the word distribution given the historical words. The observed event consist of the current word and its historical words. The current word is predicted based on the history-dependent Dirichlet parameter, which is controlled by a shared matrix $\mathbf{A}$ and the history vector $\mathbf{h}$. Different from LDA language model adaptation [5][9], LDLM directly merges the prior uncertainty of latent topics in generation of language model probability. LDLM calculates the predictive/marginal distribution over topic mixtures for Bayesian language modeling. Using LDA bigram, each document is represented as a collection of bigram events. When trigram or higher order n-gram are considered, the size of parameters $\boldsymbol{\beta}$ grows exponentially.

The total number of parameters of $\{\boldsymbol{\alpha},\boldsymbol{\beta}\}$ is $(C+C\cdot V^{n-1})$. This situation makes the implementation impractical for speech recognition. However, the proposed LDLM tackles this weakness by representing the history information in model generation procedure. We adopt the history information in generation of $\boldsymbol{\theta}$. Each history generates its own uncertainty region of topic mixtures with the shared parameters. The parameter number in $\{\mathbf{A},\boldsymbol{\beta}\}$ grows linearly as $C\cdot(n-1)V+CV$. The model complexity is controlled.

Table 1 Perplexities of LDLM with various topic sizes

	Topic Size (C)			
	C=20	C=50	C=100	C=200
LDLM	371	352	343	331
Trigram + LDLM	46.1	45.75	45.46	45.24

4. EXPERIMENTS

The Wall Street Journal (WSJ) corpus was used to evaluate the proposed LDLM for continuous speech recognition. The SI-84 training set was adopted for HMM estimation with 39 dimensional MFCC acoustic features. The HTK toolkit was used for acoustic model training and lattice generation. The '87-89 WSJ text corpus with about 30M words was used to train baseline Kneser-Ney backoff trigram and LDLM models. The LDLM was trained with 20 EM iterations. We used the 5K non-verbalized punctuation closed vocabulary for evaluation. The test set with 330 utterances was sampled from November 1992 ARPA CSR benchmark test data. We first listed the perplexities of LDLM with different size of topics in Table 1. When increasing topic size from 20 to 200, the perplexity was improved from 371 to 331. Also, the LDLM was interpolated with trigram to characterize the local lexical regularities. The perplexity of LDLM was improved to 45.24 with 200 topics. Next, we evaluated the proposed method for speech recognition. The baseline trigram was used to generate the 100-best lists. The topic-based language model by Gildea and Hofmann [4] was carried out for comparison and is denoted by GHLM. GHLM and LDLM were interpolated with baseline trigram and applied for N-best rescoring. The recognition performance is illustrated in Figure 3. We find that baseline trigram attains word error rate (WER) of 5.38%. GHLM and LDLM with 200 topics reduce WER to 5.25% and 5.19%, respectively. These results show that LDLM performs better than baseline trigram and GHLM for continuous speech recognition. However, the improvements in perplexity as well as recognition accuracy are limited because most of trigram events occur in training data. Also, we examine the error reduction rates of LDLM using various sizes of training data with 3M, 6M and 30M words and find reduction rates are 6.32%, 4.9% and 3.5%, respectively. In the case of smaller training data, LDLM performs much better due to the effectiveness of smoothing.

5. CONCLUSIONS

We presented a novel latent Dirichlet language model for continuous speech recognition. This LDLM relaxed the assumption of bag-of-words in LDA paradigm, and considered the order of historical words in the trained language model. With the superiority of LDA in topic modeling, the proposed LDLM acted as a promising topic-based language model. More interestingly, the prior uncertainty of topic mixtures of LDLM was characterized by n-gram event through the Bayesian framework. In the experiments on continuous speech recognition, we obtained desirable improvements of model perplexity and recognition accuracy. In the future, we are conducting evaluations under different cases of unseen events, and comparing LDLM with other language models.

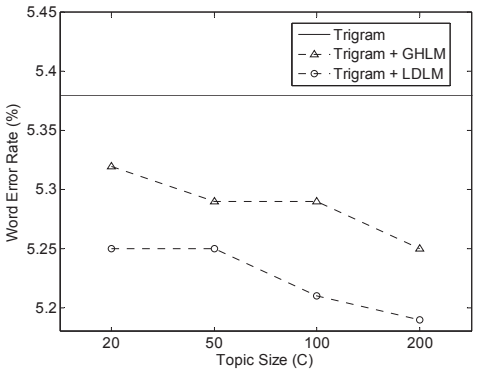

Figure 3 Word error rates (%) of GHLM and LDLM with various topic sizes

6. REFERENCES

[1] J. Bellegarda, "Exploiting latent semantic information in statistical language modeling", *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1279-1296, 2000.

[2] D. Blei, A, Ng and M. Jordan, " Latent Dirichlet allocation", *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[3] J.-T. Chien, "Association pattern language modeling", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1719-1728, 2006.

[4] D. Gildea and T. Hofmann, "Topic-based language models using EM", in *Proc. EUROSPEECH*, pp. 2167-2170, 1999.

[5] A. Heidel, H. Chang and L. Lee, "Language model adaptation using latent Dirichlet allocation and an efficient topic inference algorithm", in *Proc. INTERSPEECH*, pp. 2361-2364, 2007.

[6] T. Hofmann, "Probabilistic latent semantic indexing", in *Proc. ACM SIGIR*, pp. 50-57, 1999.

[7] M. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

[8] D. Mrva and P. C. Woodland, "A PLSA-based language model for conversational telephone speech", in *Proc. ICSLP*, pp. 2257-2260, 2004.

[9] Y. Tam and T. Schultz, "Dynamic language model adaptation using variational Bayes inference", in *Proc. INTERSPEECH*, pp. 5-8, 2005.

[10] H. Wallach, "Topic modeling: beyond bag-of-words", in *Proc. ICML*, pp. 977-984, 2006.

[11] X. Wei and W. Croft, "LDA-based document models for ad-hoc retrieval", in *Proc. ACM SIGIR*, pp. 178-185, 2006.