Using LDA to detect semantically incoherent documents

Hemant Misra and **Olivier Cappé**
LTCI/CNRS and TELECOM ParisTech
{misra,cappe}@enst.fr

François Yvon
Univ Paris-Sud 11 and LMISI-CNRS
yvon@limsi.fr

Abstract

Detecting the semantic coherence of a document is a challenging task and has several applications such as in text segmentation and categorization. This paper is an attempt to distinguish between a 'semantically coherent' true document and a 'randomly generated' false document through topic detection in the framework of latent Dirichlet analysis. Based on the premise that a true document contains only a few topics and a false document is made up of many topics, it is asserted that the entropy of the topic distribution will be lower for a true document than that for a false document. This hypothesis is tested on several false document sets generated by various methods and is found to be useful for fake content detection applications.

1 Introduction

The "Internet revolution" has dramatically increased the monetary value of higher ranking on the web search engines index, fostering the expansion of techniques, collectively known as "Web Spam", that fraudulently help to do so. Internet is indeed "polluted" with fake Web sites whose only purpose is to deceive the search engines by artificially pushing up the popularity of commercial sites, or sites promoting illegal content [1]. These fake sites are often forged using very crude content generation techniques, ranging from *web scrapping* (blending of chunks of actual contents) to simple-minded text generation techniques based on random sampling of words ("word salads"), or randomly replacing words in actual documents ("word stuffing") [2]. Among these, the latter two are easy to detect using simple statistical models of natural texts, but the former is more challenging, it being made up of actual sentences: recognizing these texts as forged requires either to resort to plagiarism detection techniques, or to automatically identify their lack of *semantic consistency*.

Detecting the consistency of texts or of text chunks has many applications in Natural Language Processing. So far, it has been used mainly in the context of automatic text segmentation, where a change in vocabulary is often the mark of topic change (Hearst, 1997), and, to a lesser extent, in discourse studies (see, e.g., (Foltz et al., 1998)). It could also serve to devise automatic metrics for text summarization or machine translation tasks.

This paper is an attempt to address the issue of differentiating between 'true' and 'false' documents on the basis of their consistency through topic modeling approach. We have used Latent Dirichlet allocation (LDA) (Blei et al., 2002) model as our main topic modeling tool. One of the aims of LDA and similar methods, including probabilistic latent semantic analysis (PLSA) (Hofmann, 2001), is to produce low dimensionality representations of texts in a "semantic space" such that most of their inherent statistical characteristics are preserved. A reduction in dimensionality facilitates storage as well as faster retrieval. Modeling discrete data has many applications in classification, categorization, topic detection, data mining, information retrieval (IR), summarization and collaborative filtering (Buntine and Jakulin, 2004).

The aim of this paper is to test LDA for establishing the semantic coherence of a document based on the premise that a real (coherent) document should discuss only a few number of topics,

[1] The annual AirWeb challenge http://airweb.cse.lehigh.edu gives a state-of-the art on current Web Spam detection techniques.

[2] The same techniques are commonly used in mail spams also.

a property hardly granted for forged documents which are often made up of random assemblage of words or sentences. As a consequence, the coherence of a document may reflect in the entropy of its posterior topic distribution or in its perplexity for the model. The entropy of the estimated topic distribution of a true document is expected to be lower than that of a fake document. Moreover, the length normalized log-likelihood of a true and coherent document may be higher as compared to that of a false and incoherent document.

In this paper, we compare two methods to estimate the posterior topic distribution of test documents, and this study is also an attempt to investigate the role of different parameters on the efficiency of these methods.

This paper is organized as follows: In Section 2, the basics of the LDA model are set. We then discuss and contrast several approaches to the problem of inferring the topic distribution of a new document in Section 3. In Section 4, we describe the corpus and experimental set-up that are used to produce the results presented in Section 5. We summarize our main findings and draw perspectives for future research in Section 6.

2 Latent Dirichlet Allocation

2.1 Basics

LDA is a probabilistic model of text data which provides a generative analog of PLSA (Blei et al., 2002), and is primarily meant to reveal hidden topics in text documents. In (Griffiths and Steyvers, 2004), the authors used LDA for identifying "hot topics" by analyzing the temporal dynamics of topics over a period of time. More recently LDA has also been used for unsupervised language model (LM) adaptation in the context of automatic speech recognition (ASR) (Hsu and Glass, 2006; Tam and Schultz, 2007; Heidel et al., 2007). Several extensions of the LDA model, such as hierarchical LDA (Blei et al., 2004), HMM-LDA (Griffiths et al., 2005), correlated topic models (Blei and Lafferty, 2005) and hidden topic Markov models (Gruber et al., 2007), have been proposed, that introduce more complex dependency patterns in the model.

Like most of the text mining techniques, LDA assumes that documents are made up of words and the ordering of the words within a document is unimportant ("bag-of-words" assumption). Contrary to the simpler Multinomial Mixture Model (see, e.g., (Nigam et al., 2000) and Section 2.4), LDA assumes that every document is represented by a topic distribution and that each topic defines an underlying distribution on words.

The generative history of a document (a bag-of-words) collection is the following: Assuming a fixed and known number of topics n_T, for each topic t, a distribution β_t over the indexing vocabulary $(w = 1 \dots n_W)$ is drawn from a Dirichlet distribution. Then, for each document d, a distribution θ_d over the topics $(t = 1 \dots n_T)$ is drawn from a Dirichlet distribution. For a document d, the document length l_d being an exogenous variable, the next step consists of drawing a topic t_i from θ_d for each position $i = 1...l_d$. Finally, a word is selected from the chosen topic t_i. Given the topic distribution, each word is thus drawn independently from every other word using a *document specific* mixture model. The probability of i^{th} word token is thus:

$$P(w_i|\theta_d, \beta) = \sum_{t=1}^{n_T} P(t_i{=}t|\theta_d) P(w_i|t_i, \beta) \quad (1)$$

$$= \sum_{t=1}^{n_T} \theta_{dt} \beta_{tw} \quad (2)$$

Conditioned on β and θ_d, the likelihood of document d is a mere product of terms such as (2), which can be rewritten as:

$$P(C_d|\theta_d, \beta) = \prod_{w=1}^{n_W} \left[\sum_{t=1}^{n_T} (\theta_{dt} \beta_{tw}) \right]^{C_{dw}} \quad (3)$$

where C_{dw} is the count of word w in d.

2.2 LDA: Training

LDA training consists of estimating the following two parameter vectors from a text collection: the topic distribution in each document d ($\theta_{dt}, t = 1...n_T, d = 1...n_D$) and the word distribution in each topic ($\beta_{tw}, t = 1...n_T, w = 1...n_W$). Both θ_d and β_t define discrete distributions, respectively over the set of topics and over the set of words. Various methods have been proposed to estimate LDA parameters, such as variational method (Blei et al., 2002), expectation propagation (Minka and Lafferty, 2002) and Gibbs sampling (Griffiths and Steyvers, 2004). In this paper, we have used the latter approach, which boils down to repeatedly going through the training data and sampling the topic assigned to each word token conditioned

on the topic assigned to all the other word tokens. Given a particular Gibbs sample, the posteriors for θ and β are [3]: Dirichlet with parameters $(K_{t1}+\lambda, \ldots, K_{tn_W}+\lambda)$ and Dirichlet with parameters $(J_{d1}+\alpha, \ldots, J_{dn_T}+\alpha)$, respectively, where K_{tw} is the number of times word w is assigned to topic t and J_{dt} is the number of times topic t is assigned to some word token in document d. Hence,

$$\beta_{tw} = \frac{K_{tw}+\lambda}{\sum_{k=1}^{n_W} K_{tk} + n_W\lambda} \quad (4)$$

$$\theta_{dt} = \frac{J_{dt}+\alpha}{\sum_{k=1}^{n_T} J_{dk} + n_T\alpha} \quad (5)$$

During the Gibbs sampling phase, β_t and θ_d are sampled from the above posteriors while the final estimates for these parameters are obtained by averaging the posterior means over the complete set of Gibbs iteration.

2.3 LDA: Testing

Training LDA model on a text collection already provides interesting insights regarding the thematic structure of the collection. This has been the primary application of LDA in (Blei et al., 2002; Griffiths and Steyvers, 2004). Even better, being a generative model, LDA can be used to make prediction regarding novel documents (assuming they use the same vocabulary as the training corpus). In a typical IR setting, where the main focus is on computing the similarity between a document d and a query d', a natural similarity measure is given by $P(C_{d'}|\theta_d, \beta)$, computed according to (3) (Buntine et al., 2004).

An alternative would be to compute the KL divergence between the topic distribution in d and d', which however requires to infer the latter quantity. As the topic distribution of a (new) document gives its representation along the latent semantic dimensions, computing this value is helpful for many applications, including text segmentation and text classification. Methods for efficiently and accurately estimating topic distribution for text documents are presented and evaluated in Section 3.

2.4 Baseline: Multinomial Mixture Model

The performance of LDA model is compared with that of the simpler multinomial mixture model (Nigam et al., 2000; Rigouste et al., 2007). In this model, every word in a document belongs to the same topic, as if the document specific topic distribution θ_d in LDA were bound to lie on one vertex of the $[0,1]^{n_T}$ simplex. Using the same notations as before (except for θ_t, which now denotes the position independent probability of topic t in the collection), the probability of a document is:

$$P(C_d|\theta_t, \beta) = \sum_{t=1}^{n_T} \theta_t \prod_{w=1}^{n_W} \beta_{tw}^{C_{dw}} \quad (6)$$

This model can be trained through expectation maximization (EM), using the following reestimation formulas, where (7) defines the E-step; (8) and (9) define the M-step.

$$P(t|C_d, \theta, \beta) = \frac{\theta_t \prod_{w=1}^{n_W} (\beta'_{tw})^{C_{dw}}}{\sum_{t=1}^{n_T} \theta_t \prod_{w=1}^{n_W} (\beta_{tw})^{C_{dw}}} \quad (7)$$

$$\theta'_t \propto \alpha + \sum_{d=1}^{n_D} P(t|C_d, \theta, \beta) \quad (8)$$

$$\beta'_{tw} \propto \lambda + \sum_{d=1}^{n_D} C_{dw} P(t|C_d, \theta, \beta) \quad (9)$$

As suggested in (Rigouste et al., 2007), we initialize the EM algorithm by drawing initial topic distributions from a prior Dirichlet distribution with hyper-parameter $\alpha = 1$. $\beta = 0.1$ in all the experiments.

During testing, the parameters of the multinomial models are used to estimate the posterior topic distribution in each document using (7). The likelihood of a test document is given by (6).

3 Inferring the Topic Distribution of Test Documents

$P(C_d|\theta_d)$, the conditional probability of a document d given θ_d is obtained using (3) [4]. Computing the likelihood of a test document requires to integrate this quantity over θ; likewise for the computation of the posterior distribution of θ. This integral has no close form solution, but can be approximated using Monte-Carlo sampling techniques as:

$$P(C_d) \approx \frac{1}{M} \sum_{m=1}^{M} P(C_d|\theta^{(m)}) \quad (10)$$

where $\theta^{(m)}$ denotes the m^{th} sample from the Dirichlet prior, and M is the number of Monte

[3] assuming non-informative priors with hyper-parameters α and λ for the Dirichlet distribution over topics and the Dirichlet distribution over words respectively

[4] The dependence on β is dropped for simplicity. β is learned during training and kept fixed during testing.

Carlo samples. Given the typical length of documents and the large vocabulary size, small scale experiments convinced us that a cruder approximation was in order, as the sum in (10) is dominated by the maximum value. We thus contend ourselves to solve:

$$\theta^* = \underset{\theta, \sum_t \theta_t = 1}{\operatorname{argmax}} P(C_d|\theta) \tag{11}$$

and use this value to approximate $P(C_d)$ using (3).

The maximization program (11) has no close form solution. However, the objective function is differentiable and log-concave, and can be optimized in a number of ways. We considered two different algorithms: an EM-like approach, initially introduced in (Heidel et al., 2007), and an exponentiated gradient approach (Kivinen and Warmuth, 1997; Globerson et al., 2007).

The first approach implements an iterative procedure based on the following update rule:

$$\theta_{dt} \leftarrow \frac{1}{l_d} \sum_{w=1}^{n_W} \frac{C_{dw}\theta_{dt}\beta_{tw}}{\sum_{t'=1}^{n_T} \theta_{dt'}\beta_{t'w}} \tag{12}$$

Although no justification was given in (Heidel et al., 2007), it can be shown that this update rule converges towards a global optimum of the likelihood. Let θ and θ' be two topic distributions in the n_T-dimensional simplex, $L(\theta) = \log P(C_d|\theta)$, and $\rho_t(w,\theta) = \frac{\theta_t \beta_{tw}}{\sum_{t'} \theta_{t'}\beta_{t'w}}$. We define an auxiliary function $Q(\theta,\theta') = \sum_w C_w(\sum_t \rho_t(w,\theta)\log(\theta'_t))$. $Q(\theta,\theta')$ is concave in θ', and performs the role played by the auxiliary function in the EM algorithm. Simple calculus suffices to prove that (i) the update (12) maximizes in θ' the function $Q(\theta,\theta')$, and (ii) $Q(\theta,\theta') - Q(\theta,\theta) \geq L(\theta') - L(\theta)$, which stems from the concavity of the $\log$. At an optimum of $Q(\theta,\theta')$ the positivity of the first term implies the positivity of the second. Maximizing Q using the update rule (12) thus increases the likelihood and repeating this update converges towards the optimum value. We experimented both with an unsmoothed (12) and with a smoothed version of this update rule. The unsmoothed version yielded a slightly better result than the smoothed one.

Exponentiated gradient (Kivinen and Warmuth, 1997; Globerson et al., 2007) yields an alternative update rule:

$$\theta_{dt} \leftarrow \theta_{dt} \exp\left(\eta \sum_{w=1}^{n_W} \frac{C_{dw}\beta_{tw}}{\sum_{t'=1}^{n_T} \theta_{dt'}\beta_{t'w}}\right) \tag{13}$$

where η defines the convergence rate. In this form, the update rule does not preserve the normalization of θ, which needs to be performed after every iteration.

A systematic comparison of these rules was carried out, yielding the following conclusions:

- the convergence of the EM-like method is very fast. Typically, it requires less than half a dozen iterations to converge. After convergence, the topic distribution estimated by this method for a subset of train documents was always very close (as measured by the KL-divergence) to the respective topic distribution of the same documents observed at the end of the LDA training. Taking $n_T = 50$, the average KL divergence for a set of 4,500 documents was found to be less than 0.5.

- exponentiated gradient has a more erratic behaviour, and requires a careful tuning of η on a *per document basis*. For large values of η, the update rule (13) sometimes fails to converge; smaller values of η allowed to consistently reach convergence, but required more iterations (typically 20-30). On a positive side, on an average, the topic distributions estimated by this method are better than the ones obtained with the EM-like algorithm.

Based on these findings, we decided to use the EM-like algorithm in all our subsequent experiments.

4 Experimental protocol

4.1 Training and test corpora

The Reuters Corpus Volume 1 (RCV1) (Lewis et al., 2004) is a collection of over 800,000 news items in English from August 1996 to August 1997. Out of the entire RCV1 dataset, we selected 27,672 documents (news items) for training (**TrainReuters**) and 23,326 documents for testing (**TestReuters**). The first 4000 documents from the **TestReuters** dataset were used as true documents (**TrueReuters**) in the experiments reported in this paper. The vocabulary size in the train set, after removing the function words, is $93,214$.

Along with these datasets of "true" documents, three datasets of fake documents were also created. Document generation techniques are many: here we consider documents made by mixing short passages from various texts and documents made

by assembling randomly chosen words (sometimes called as "word salads"). In addition, we also consider the case of documents generated with a stochastic language model (LM). Our "fake" test documents are thus composed of:

- (**SentenceSalad**) obtained by randomly picking sentences from **TestReuters**.
- (**WordSalad**) created by generating random sentences from a conventional unigram LM trained on **TrainReuters**.
- (**Markovian**) created by generating random sentences from a conventional 3-gram LM trained on **TrainReuters**.

Each of these forged document set contains 4,000 documents.

To assess the performance on out-of-domain data, we replicated the same tests using 2,000 Medline abstracts (Ohta et al., 2002). 1,500 documents were used either to generate fake documents by picking sentences randomly or to train an LM and then using the LM to generate fake documents. The remaining 500 abstracts were set aside as "true" documents (**TrueMedline**).

4.2 Performance Measurements : EER

The entropy of the topic distribution is computed as $H = -\sum_{j=1}^{T} \hat{\theta}_{dj} \log \hat{\theta}_{dj}$. The other measure of interest is the average 'log-likelihood per word' (LLPW) [5].

While evaluating the performance of our system, two types of errors are encountered: false acceptance (FA) when a false document is accepted as a true document and false rejection (FR) when a true document is rejected as a false document. The rate of FA and FR is dependent on the threshold used for taking the decision, and usually the performance of a system is shown by its receiver operating characteristic (ROC) curve which is a plot between FA and FR rates for different values of threshold. Instead of reporting the performance of a system based on two error rates (FA and FR), the general practice is to report the performance in terms of equal-error-rate (EER). The EER is the error rate at the threshold where FA rate = FR rate.

In our system, a threshold on entropy (or LLPW) is used for taking the decision, and all the documents having their entropy (or LLPW) below (or above) the threshold are accepted as true documents. The EER is obtained on the test set by changing the threshold on the test set itself, and the best results thus obtained are reported.

[5] This measure is directly related to the text perplexity in the model, according to perplexity $= 2^{-\text{average log-likelihood per word}}$

5 Detecting semantic inconsistency

5.1 Detecting fake documents with LDA and Multinomial mixtures

In the first set of experiments, the LLPW and entropy of the topic distribution (the two measures) of the Multinomial mixture and LDA models were compared to check the ability of these two measures and models in discriminating between true and false documents. These results are summarized in Table 1.

TrueReuters vs.	**Multinomial**	
	LLPW	Entropy
SentenceSalad	15.3%	48.8%
WordSalad	9.3%	35.8%
Markovian	17.6%	38.9%
TrueReuters vs.	**LDA**	
	LLPW	Entropy
SentenceSalad	18.9%	0.88%
WordSalad	9.9%	0.13%
Markovian	25.0%	0.28%

Table 1: *Performance of the Multinomial Mixture and LDA*

For the multinomial mixture model, the LLPW measure is able to discriminate between true and false documents to a certain extent. As expected (not shown here), the LLPW of the true documents is usually higher than that of the false documents. In contrast, the entropy of the posterior topic distribution does not help much in discriminating between true and false documents. In fact it remains close to zero (meaning that only one topic is "active") both for true and false documents.

The behaviour of the LDA scores is entirely different. The perplexity scores (LLPW) of true and fake texts are comparable, and do not make useful predictors. In contrast, the entropy of the topic distribution allows to sort true documents from fake ones with a very high accuracy for all kinds of fake texts considered in this paper. Both results stem from the ability of LDA to assign a different topic to each word occurrence.

Similar pattern is observed for our three false test sets (against the TrueReuters set) with small

variations The texts generated with a Markov model, no matter the order, have the highest entropy, reflecting the absence of long range correlation in the generation model. Though the texts generated by mixing sentences are more confusing with the true documents, the performance is still less than 1% EER. Texts mixing a high number of topics (e.g., Sentence Salads) are almost as likely as natural texts that address only a few topics. However, the former has much higher entropy of the topic distribution due to a large number of topics being active in such texts (see also Figure 1).

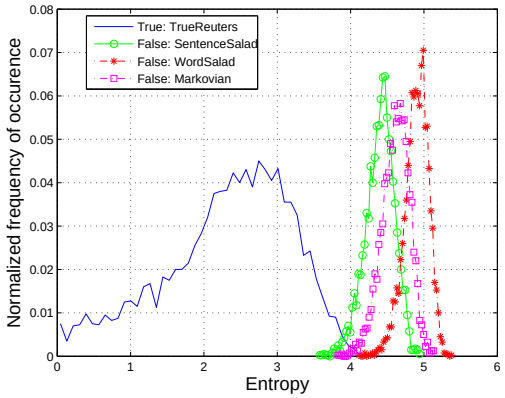

Figure 1: Histogram of entropy of θ for different true and false document sets.

It is noteworthy that both the predictors (LLPW and Entropy) give complementary clues regarding a text category. A linear combination of these two scores (the weight to the LLPW score is 0.1) allows to substantially improve over these baseline results, yielding a relative improvement (in EER) of +20.0% for the sentence salads, +20.8% for the word salads, and +27.3% for the Markov Models.

5.2 Effect of the number of topics

In this part, we investigate the performance of LDA in detecting false documents when the number of topics is changed. Increasing the number of topics means higher memory requirements both during training and testing. Though the results are shown only for SentenceSalad, similar trend is observed for WordSalad and Markovian.

The numbers in Table 2 show that the performance obtained with the LLPW score consistently improve with an increase in the number of topics, though the % improvement obtained when the number of topics exceeds 200 is marginal. In contrast, the best performance in case of entropy is achieved at 50 topics and slowly degrades when a more complex model is used.

Number of Topics	LLPW	Entropy
10	27.9	1.88
50	18.9	0.88
100	16.0	0.93
200	14.8	0.90
300	13.8	1.05
400	13.6	1.10

Table 2: *EER from LLPW and Entropy distribution for TrueReuters against SentenceSalad.*

5.3 Detecting "noisy" documents

In this section, we study fake documents produced by randomly changing words in true documents (the TrueReuters dataset). In each document, a fixed percentage of content words is randomly replaced by any other word from the training vocabulary [6]. This percentage was varied from 5 to 100 and EER for these corrupted document sets is computed at each % corruption level (Figure 2). As

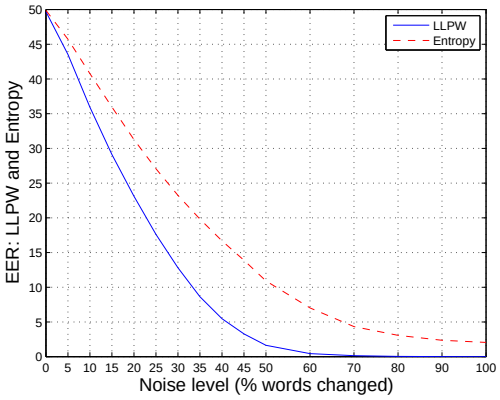

Figure 2: *EER at various noise levels*

expected, the EER is very high at low noise levels, and as the noise level is increased, EER gets lower. When only a few words are changed in a true document, it retrains the properties of a true document (high LLPW and low entropy). However, as more number of words are changed in a true document,

[6] When the replacement words are chosen from a small set of very specifi c words, the fake document generation strategy is termed as "word stuffi ng".

it starts showing the characteristics of a false document (low LLPW and high entropy). These results suggest that our semantic consistency tests are too crude a measure to detect a small number of inconsistencies, such as the ones found in the state-of-the-art OCR or ASR systems' outputs. On the other hand, it confirms the numerous studies that have shown that topic detection (and topic adaptation) or text categorization tasks can be performed with the same accuracy for moderately noisy texts and clean texts, a finding which warrants the topic-based LM adaptation strategies deployed in (Heidel et al., 2007; Tam and Schultz, 2007).

The difference in the behavior of our two predictors is striking. The EER obtained using LLPW drops more quickly than the one obtained with entropy of the topic distribution. It suggests that the influence of "corrupting" content words (mostly with low β_{tw}) is heavy on the LLPW, but the topic information is not lost till a majority of the "uncorrupted" content words belong to the same topic.

5.4 Effect of the document length

In this section, we study the robustness of our two predictors with respect to the document length by progressively increasing the number of content words in a document (true or fake). As can be seen from Figure 3, the entropy of the posterior topic distribution starts to provide a reasonable discrimination (5% EER) when the test documents contain about 80 to 100 content words, and attains results comparable to those reported earlier in this paper when this number doubles. This definitely rules out this method as a predictor of the semantic consistency of a sentence: we need to consider at least a paragraph to get acceptable results.

5.5 Testing with out-of-domain data

In this section, we study the robustness of our predictors on out-of-domain data using a small excerpt of abstracts from the Medline database. Both true and fake documents are from this dataset. The results are summarized in Table 3. The performance on out-of-domain documents is poor,

TrueMedline vs.	LLPW	Entropy
SentenceSalad	31.23%	22.13%
WordSalad	30.03%	19.46%
Markovian	36.51%	23.63%

Table 3: *Performance of LDA on PubMed abstracts*

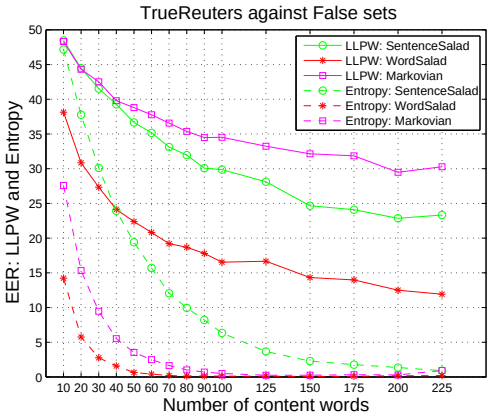

Figure 3: *EER with change in number of content words used for LDA analysis. EER based on: LLPW of TrueReuters and false document sets (solid line) and Entropy of topic distribution of TrueReuters and false document sets (dashed line).*

though the entropy of the topic distribution is still the best predictor. The reasons for this failure are obvious: a majority of the words occurring in these documents (true or fake) are, from the perspective of the model, characteristic of one single Reuters topic (health and medicine). They cannot be distinguished either in terms of perplexity or in terms of topic distribution (the entropy is low for all the documents). It is interesting to note that all the out-of-domain Medline data can be separated from the in-domain TrueReuters data with good accuracy on the basis of the lower LLPW of the former as compared to the higher LLPW of the latter.

6 Conclusion

In the LDA framework, this paper investigated two methods to infer the topic distribution in a test document. Further, the paper suggested that the coherence of a document can be evaluated based on its topic distribution and average LLPW, and these measures can help to discriminate between true and false documents. Indeed, through experimental results, it was shown that entropy of the topic distribution is lower and average LLPW of true documents is higher for true documents and the former measure was found to be more effective. However, the poor performance of this method on out-of-domain data suggests that we need to use a much larger training corpus to build a robust fake document detector. This raises the issue of train-

ing LDA model with very large collections. In future we would like to explore the potential of this method for text segmentation tasks.

Acknowledgment

This research was supported by the European Commission under the contract *FP6-027026-K-Space*. The views expressed in this paper are those of the authors and do not necessarily represent the views of the commission.

References

Blei, David and John Lafferty. 2005. Correlated topic models. In *Advances in Neural Information Processing Systems (NIPS'18)*, Vancouver, Canada.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2002. Latent Dirichlet allocation. In Dietterich, Thomas G., Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 14, pages 601–608, Cambridge, MA. MIT Press.

Blei, David M., Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. 2004. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems (NIPS)*, volume 16, Vancouver, Canada.

Buntine, Wray and Aleks Jakulin. 2004. Applying discrete PCA in data analysis. In Chickering, M. and J. Halpern, editors, *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI'04)*, pages 59–66. AUAI Press 2004.

Buntine, Wray, Jaakko Löfström, Jukka Perkiö, Sami Perttu, Vladimir Poroshin, Tomi Silander, Henry Tirri, Antti Tuominen, and Ville Tuulos. 2004. A scalable topic-based open source search engine. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 228–234, Beijing, China.

Foltz, P.W., W. Kintsch, and T.K. Landauer. 1998. The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, 25(2-3):285–307.

Globerson, Amir, Terry Y. Koo, Xavier Carreras, and Michael Collins. 2007. Exponentiated gradient algorithms for log-linear structured prediction. In *International Conference on Machine Learning*, Corvallis, Oregon.

Griffiths, Thomas L. and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 (supl 1):5228–5235.

Griffiths, Thomas L., Mark Steyvers, David M. Blei, and Joshua Tenenbaum. 2005. Integrating topics and syntax. In *Proceedings of NIPS, 17*, Vancouver, CA.

Gruber, Amit, Michal Rosen-Zvi, and Yair Weiss. 2007. Hidden topic Markov models. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, San Juan, Puerto Rico, March.

Hearst, Marti. 1997. TextTiling: Segmenting texts into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Heidel, Aaron, Hung an Chang, and Lin shan Lee. 2007. Language model adaptation using latent Dirichlet allocation and an efficient topic inference algorithm. In *Proceedings of European Conference on Speech Communication and Technology*, Antwerp, Belgium.

Hofmann, Thomas. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal*, 42(1):177–196.

Hsu, Bo-June (Paul) and Jim Glass. 2006. Style & topic language model adaptation using HMM-LDA. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

Kivinen, Jyrki and Manfrud K. Warmuth. 1997. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132:1–63.

Lewis, David D., Yiming Yang, Tony Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *Machine Learning Research*, 5:361–397.

Minka, Thomas and John Lafferty. 2002. Expectation-propagation for the generative aspect model. In *Proceedings of the conference on Uncertainty in Artificial Intelligence (UAI)*.

Nigam, K., A. K. McCallum, S. Thrun, and T. M. Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134.

Ohta, Tomoko, Yuka Tateisi, Hideki Mima, Jun ichi Tsujii, and Jin-Dong Kim. 2002. The GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of Human Language Technology Conference*, pages 73–77.

Rigouste, Loïs, Olivier Cappé, and François Yvon. 2007. Inference and evaluation of the multinomial mixture model for text clustering. *Information Processing and Management*, 43(5):1260–1280, September.

Tam, Yik-Cheung and Tanja Schultz. 2007. Correlated latent semantic model for unsupervised LM adaptation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Honolulu, Hawaii, U.S.A.