Term Weighting Schemes for Latent Dirichlet Allocation

Andrew T. Wilson
Sandia National Laboratories
PO Box 5800, MS 1323
Albuquerque, NM 87185-1323, USA
atwilso@sandia.gov

Peter A. Chew
Moss Adams LLP
6100 Uptown Blvd. NE, Suite 400
Albuquerque, NM 87110-4489, USA
Peter.Chew@MossAdams.com

Abstract

Many implementations of Latent Dirichlet Allocation (LDA), including those described in Blei et al. (2003), rely at some point on the removal of stopwords, words which are assumed to contribute little to the meaning of the text. This step is considered necessary because otherwise high-frequency words tend to end up scattered across many of the latent topics without much rhyme or reason. We show, however, that the 'problem' of high-frequency words can be dealt with more elegantly, and in a way that to our knowledge has not been considered in LDA, through the use of appropriate weighting schemes comparable to those sometimes used in Latent Semantic Indexing (LSI). Our proposed weighting methods not only make theoretical sense, but can also be shown to improve precision significantly on a non-trivial cross-language retrieval task.

1 Introduction

Latent Dirichlet Allocation (LDA) (Blei et al., 2003), like its more established competitors Latent Semantic Indexing (LSI) (Deerwester et al., 1990) and Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999), is a model which is applicable to the analysis of text corpora. It is claimed to differ from LSI in that LDA is a generative Bayesian model (Blei et al., 2003), although this may depend upon the manner in which one approaches LSI (see for example Chew et al. (2010)). In LDA as applied to text analysis, each document in the corpus is modeled as a mixture over an underlying set of topics, and each topic is modeled as a probability distribution over the terms in the vocabulary.

As the newest among the above-mentioned techniques, LDA is still in a relatively early stage of development. It is also sufficiently different from LSI, probably the most popular and well-known compression technique for information retrieval (IR), that many practitioners of LSI may perceive a 'barrier to entry' to LDA. This in turn perhaps explains why notions such as term weighting, which have been commonplace in LSI for some time (Dumais, 1991), have not yet found a place in LDA. In fact, it is often assumed that weighting is unnecessary in LDA. For example, Blei et al. (2003) contrast the use of tf-idf weighting in both non-reduced space (Salton and McGill, 1983) and LSI on the one hand with PLSI and LDA on the other, where no mention is made of weighting. Ramage et al. (2008) propose a simple term-frequency weighting scheme for tagged documents within the framework of LDA, although term weighting is not their focus and their scheme is intended to incorporate document tags into the same model that represents the documents themselves.

In this paper, we produce evidence that term weighting should be given consideration within LDA. First and foremost, this is shown empirically through a non-trivial multilingual retrieval task which has previously been used as the basis for tests of variants of LSI. We also show that term weighting allows one to avoid maintenance of stoplists, which can be awkward especially for multilingual data. With appropriate term weighting, high-frequency words (which might otherwise be eliminated as stopwords) are assigned naturally to topics

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 465–473,
Los Angeles, California, June 2010.

by LDA, rather than dominating and being scattered across many topics as happens with the standard uniform weighting. Our approach belies the usually unstated, but widespread, assumption in papers on LDA that the removal of stopwords is a necessary pre-processing step (see e.g. Blei et al. (2003); Griffiths and Steyvers (2004)).

It might seem that to demonstrate this it would be necessary to perform a test that directly compares the results when stoplists are used to those when weighting are used. However, we believe that stopwords are highly ad-hoc to begin with. Assuming a vocabulary of n words and a stoplist of x items, there are (at least in theory) $\binom{n}{x}$ possible stoplists. To be sure that *no* stoplist improves on a particular term weighting scheme we would have to test every one of these. In addition, our tests are with a multilingual dataset, which raises the issue that a domain-appropriate stoplist for a particular corpus and language may not be available. This is even more true if we pre-process the dataset morphologically (for example, with stemming). Therefore, rather than attempting a direct comparison of this type, we take the position that it is possible to sidestep the need for stoplists and to do so in a non-ad-hoc way.

The paper is organized as follows. Section 2 describes the general framework of LDA, which has only very recently been applied to cross-language IR. In Section 3, we look at alternatives to the 'standard' uniform weighting scheme (i.e., lack of weighting scheme) commonly used in LDA. Section 4 discusses the framework we use for empirical testing of our hypothesis that a weighting scheme would be beneficial. We present the results of this comparison in Section 5 along with an impressionistic comparison of the output of the different alternatives. We conclude in Section 6.

2 Latent Dirichlet Allocation

Our IR framework is multilingual Latent Dirichlet Allocation (LDA), first proposed by Blei et al. (2003) as a general Bayesian framework with initial application to topic modeling. It is only very recently that variants of LDA have been applied to cross-language IR: examples are Cimiano et al. (2009) and Ni et al. (2009).

As an approach to topic modeling, LDA relies on the idea that the tokens in a document are drawn independently from a set of topics where each topic is a distribution over types (words) in the vocabulary. The mixing coefficients for topics within each document and weights for types in each topic can be specified *a priori* or learned from a training corpus. Blei et al. initially proposed a variational model (2003) for learning topics from data. Griffiths and Steyvers (2004) later developed a Markov chain Monte Carlo approach based on collapsed Gibbs sampling.

In this model, the mixing weights for topics within each document and the multinomial coefficients for terms within each topic are hidden (latent) and must be learned from a training corpus. Blei et al. (2003) proposed LDA as a general Bayesian framework and gave a variational model for learning topics from data. Griffiths and Steyvers (2004) subsequently developed a stochastic learning algorithm based on collapsed Gibbs sampling. In this paper we will focus on the Gibbs sampling approach.

2.1 Generative Document Model

The LDA algorithm models the D documents in a corpus as mixtures of K topics where each topic is in turn a distribution over W terms. Given $\boldsymbol{\theta}$, the matrix of mixing weights for topics within each document, and $\boldsymbol{\phi}$, the matrix of multinomial coefficients for each topic, we can use this formulation to describe a generative model for documents (Alg. 1).

Restating the LDA model in linear-algebraic terms, we can say that the product of $\boldsymbol{\phi}$ (the $K \times W$ column-stochastic topic-by-type matrix) and $\boldsymbol{\theta}$ (the $D \times K$ column-stochastic topic-by-document matrix) is the original $D \times W$ term-by-document matrix. In this sense, LDA computes a matrix factorization of the term-by-document matrix in the same way that LSI or non-negative matrix factorization (NMF) do. In fact, LDA is a special case of NMF, but unlike in NMF, there is a unique factorization in LDA. We see this as a feature recommending LDA above NMF.

Our objective is to reverse the generative model to learn the contents of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ given a training corpus D, a number of topics K, and symmetric Dirichlet prior distributions over both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ with hyperparameters α and β, respectively.

for $k = 1$ to K **do**
  Draw $\phi^k \sim \text{Dirichlet}(\beta)$
end for
for $d = 1$ to D **do**
  Draw $\theta \sim \text{Dirichlet}(\alpha)$
  Draw $N \sim \text{Poisson}(\xi)$
  for $i = 1$ to N **do**
    Draw $z \sim \text{Multinomial}(\theta)$
    Draw $w \sim \text{Multinomial}(\phi^{(z)})$
  end for
end for

Algorithm 1: Generative algorithm for LDA. This will generate D documents with N tokens each. Each token is drawn from one of K topics. The distributions over topics and terms have Dirichlet hyperparameters α and β respectively. The Poisson distribution over the token count may be replaced with any other convenient distribution.

2.2 Learning Topics via Collapsed Gibbs Sampling

Rather than learn $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ directly, we use collapsed Gibbs sampling (Geman et al. (1993), Chatterji and Pachter (2004)) to learn the latent assignment of tokens to topics $\mathbf{z}$ given the observed tokens $\mathbf{x}$.

The algorithm operates by repeatedly sampling each z_{ij} from a distribution conditioned on the values of all other elements of $\mathbf{z}$. This requires maintaining counts of tokens assigned to topics globally and within each document. We use the following notation for these sums:

N_{ijk}**:** Number of tokens of type w_i in document d_j assigned to topic k

N_{ijk}^{-st}**:** The sum N_{ijk} with the contribution of token x_{st} excluded

We indicate summation over all values of an index with $(\cdot)$.

Given the current state of $\mathbf{z}$ the conditional probability of z_{ij} is:

$$p(z_{ij} = k|\mathbf{z}^{-ij}, \mathbf{x}, d, \alpha, \beta) = \\ p(x_{ij}|\phi^k) \quad p(k|d_j) \quad \propto \\ \frac{N_{i(\cdot)k}^{-ij} + \beta}{N_{(\cdot)(\cdot)k}^{-ij} + W\beta} \frac{N_{(\cdot)jk}^{-ij} + \alpha}{N_{(\cdot)j(\cdot)} + T\alpha} \tag{1}$$

As Griffiths and Steyvers (2004) point out, this is an intuitive result. The first term, $p(x_{ij}|\phi^k)$, indicates the importance of term x_{ij} in topic k. The second term, $p(k|d_j)$, indicates the importance of topic k in document j. The sum of the terms is normalized implicitly to 1 when we draw each new z_{ij}.

We sample a new value for z_{ij} for every token x_{ij} during each iteration of Gibbs sampling. We run the sampler for a burn-in period of a few hundred iterations to allow it to reach its converged state and then estimate $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ from $\mathbf{z}$ as follows:

$$\theta_{jk} = \frac{N_{(\cdot)jk} + \alpha}{N_{(\cdot)j(\cdot)} + T\alpha} \tag{2}$$

$$\phi_{ki} = \frac{N_{i(\cdot)k} + \beta}{N_{(\cdot)(\cdot)k} + W\beta} \tag{3}$$

2.3 Classifying New Documents

In LSI, new documents not in the original training set can be 'projected' into the semantic space of the training set. The equivalent process in LDA is one of *classification*: given a corpus D' of one or more new documents we use the existing topics $\boldsymbol{\phi}$ to compute a maximum a posteriori estimate of the mixing coefficients $\boldsymbol{\theta}'$. This follows the same Monte Carlo process of repeatedly resampling a set of token-to-topic assignments $\mathbf{z}'$ for the tokens $\mathbf{x}'$ in the new documents. These new tokens are used to compute the first term $p(k|d_j)$ in Eq. 1. We re-use the topic assignments $\mathbf{z}$ from the training corpus to compute the second term $p(x_{ij}|\phi_k)$. Tokens with new types that were not present in the vocabulary of the training corpus do not participate in classification.

The resulting distribution $\boldsymbol{\theta}'$ essentially encodes how likely each new document is to relate to each of the K topics. We can use this matrix to compute pairwise similarities between any two documents from either corpus (training or newly-classified). Whereas in LSI it may make sense to compute similarity between documents using the cosine metric (since the 'dimensions' defining the space are orthogonal), we compute similarities in LDA using either the symmetrized Kullback-Leibler (KL) or Jensen-Shannon (JS) divergences (Kullback and Leibler (1951), Lin (2002)) since these are methods of measuring the similarity between probability distributions.

3 Term Weighting Schemes and LDA

The standard approach presented above assumes, effectively, that each token is equally important in calculating the conditional probabilities. From both an information-theoretic and a linguistic point of view, however, it is clear that this is not the case. In English, a term such as 'the' which occurs with high frequency in many documents does not contribute as much to the meaning of each document as a lower-frequency term such as 'corpus'. It is an axiom of information theory that an event a's information content (in bits) is equal to $\log_2 \frac{1}{p(a)} = -\log_2 p(a)$.

Treating tokens as events, we can say that the information content of a particular token of type t is $-\log_2 p(t)$. Furthermore, as is well-known, we can estimate $p(t)$ from observed frequencies in a corpus: it is simply the number of tokens of type t in the corpus, divided by the total number of tokens in the corpus. For high-probability terms such as 'the', therefore, $-\log_2 p(t)$ is low. Our basic hypothesis is that recalculating $p(z_{ij}|\mathbf{z}, \mathbf{x}, \alpha, \beta)$ to take the information content of each token into account will improve the results of LDA. Specifically, we have incorporated a weighting term into Eq. 1 by replacing the counts denoted N with weights denoted M.

$$p(z_{ij} = k|\mathbf{z}^{-ij}, \mathbf{x}, d, \alpha, \beta) \propto \frac{M^{-ij}_{i(\cdot)k} + \beta}{M^{-ij}_{(\cdot)(\cdot)k} + W\beta} \frac{M^{-ij}_{(\cdot)jk} + \alpha}{M_{(\cdot)j(\cdot)} + T\alpha} \quad (4)$$

Here M_{ijk} is the total *weight* of tokens of type i in document j assigned to topic k instead of the total *number* of tokens. All of the machinery for Gibbs sampling and the estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ from $\mathbf{z}$ remains unchanged.

We appeal to an urn model to explain the intuition behind this approach. In the original LDA formulation, each topic ϕ can be modeled as an urn containing a large number of balls of uniform size. Each ball assumes one of W different colors (one color for each term in the vocabulary). The frequency of occurrence of each color in the urn is proportional to the corresponding term's weight in topic ϕ. We incorporate a term weighting scheme by making the size of each ball proportional to the weight of its corresponding term. This makes the probability of drawing the ball for a term w proportional to both the term weight $m(w)$ and its multinomial weight ϕ^w:

$$p(w|\phi, \beta, m) = \frac{\phi^w \, m(w)}{\sum_{w \in W} m(w)} \quad (5)$$

We can now expand Eq. 4 to obtain a new sampling equation for use with the Gibbs sampler.

$$p(z_{ij} = k|\mathbf{z}^{-ij}, \mathbf{x}, \mathbf{d}, m, \alpha, \beta) = \frac{m(x_i)N^{-ij}_{i(\cdot)k} + \beta}{\sum_w m(w)N^{-ij}_{w(\cdot)k} + W\beta} \frac{\sum_w m(w)N^{-ij}_{wjk} + \alpha}{\sum_w m(w)N_{wj(\cdot)} + T\alpha} \quad (6)$$

If all weights $m(w) = 1$ this reduces immediately to the standard LDA formulation in Eq. 1.

The information measure we describe above is constant for a particular term across the entire corpus, but it is possible to conceive of other, more sophisticated weighting schemes as well, for example those where term weights vary by document. Pointwise mutual information (PMI) is one such weighting scheme which has a solid basis in information theory and has been shown to work well in the context of LSI (Chew et al., 2010). According to PMI, the weight of a given term w in a given document d is the pointwise mutual information of the term and document, or $-\log_2 \frac{p(w|d)}{p(w)}$. Extending the LDA model to accommodate PMI is straightforward. We replace $m(x_i)$ and $m(w)$ in Eq. 4 with $m(x_i, d)$ as follows.

$$\begin{aligned} m(x_i, d) &= -\log_2 \frac{p(x_i|d)}{p(x_i)} \\ &= -\log_2 \frac{\#[\text{tokens of type } x_i \text{ in } d]}{\#[\text{tokens of type } x_i]} \end{aligned} \quad (7)$$

It is possible for PMI of a term within a document to be negative. When this happens, we clamp the weight of the offending term to zero in that document. In practice, we observe this only with common words (e.g. 'and', 'in', 'of', 'that', 'the' and 'to' in English) that are assigned very low weight everywhere else in the corpus. This clamping does not noticeably affect the results.

In the next sections, we describe tests which have enabled us to evaluate empirically which of these formulations works best in practice.

4 Testing Framework

In this paper, we chose to test our hypotheses with the same cross-language retrieval task used in a number of previous studies of LSI (e.g. Chew and Abdelali (2007)). Briefly, the task is to train an IR model on one particular multilingual corpus, then deploy it on a separate multilingual corpus, using a document in one language to retrieve related documents in other languages. This task is difficult because of the size of the datasets involved. Its usefulness becomes apparent when we consider the following two use cases: a human wishing (1) to use a search engine to retrieve relevant documents in many languages regardless of the language in which the query is posed; or (2) to produce a clustering or visualization of documents according to their topics even when the documents are in different languages.

The training corpus consists of the text of the Bible in 31,226 parallel chunks, corresponding generally to verses, in Arabic, English, French, Russian and Spanish. These data were obtained from the Unbound Bible project (Biola University (2006)). The test data, obtained from `http://www.kuran.gen.tr/`, is the text of the Quran in the same 5 languages, in 114 parallel chunks corresponding to suras (chapters). The task, in short, is to use the training data to inform whatever linguistic, semantic, or statistical model is being tested, and then to infer characteristics of the test data in such a way that the test documents can automatically be matched with their translations in other languages. Though the documents come from a specific domain (scriptural texts), what is of interest is comparative results using different weighting schemes, holding the datasets and other settings constant. The training and test datasets are large enough to allow statistically significant observations to be made, and if a significant difference is observed between experiments using two settings, it is to be expected that similar basic differences would be observed with any other set of training and test data. In any case, it should be noted that the Bible and Quran were written centuries apart, and in different original languages; we believe this contributes to a clean separation of training and test data, and makes for a non-trivial retrieval task.

In our framework, a term-by-document matrix is formed from the Bible as a parallel verse-aligned corpus. We employed two different approaches to tokenization, one (word-based tokenization) in which text was tokenized at every non-word character, and the other (unsupervised morpheme-based tokenization) in which after word-based tokenization, a further pre-processing step (based on Goldsmith (2001)) was performed to add extra breaks at every morpheme. It is shown elsewhere (Chew et al., 2010) that this step leads to improved performance with LSI. In each verse, all languages are concatenated together, allowing terms (either morphemes or words) from all languages to be represented in every verse. Cross-language homographs such as 'mien' in English and French are treated as distinct terms in our framework. Thus, if there are L languages, D documents (each of which is translated into each of the L languages), and W distinct linguistic terms across all languages, then the term-by-document matrix is of dimensions W by D (not W by $D \times L$); with the Bible as a training corpus, the actual numbers in our case are 160,345 $\times$ 31,226. As described in Sec. 2.2, we use this matrix as the input to a collapsed Gibbs sampling algorithm to learn the latent assignment of tokens in all five languages to language-independent topics, as well as the latent assignment of language-independent topics to the multilingual (parallel) documents. In general, we specified, arbitrarily but consistently across all tests, that the number of topics to be learned should be 200. Other parameters for the Gibbs sampler held constant were the number of iterations for burn-in (200) and the number of iterations for sampling (1).

To evaluate our different approaches to weighting, we use classification as described in Sec. 2.3 to obtain, for each document from the Quran test corpus, a probability distribution across the topics learned from the Bible. While in training we have D multilingual documents, in testing we have $D' \times L$ documents, each in a specific language, for which a distribution is computed. For the Quran data, this amounts to 114 $\times$ 5 = 570 documents. This is because our goal is to match documents with their translations in other languages using just the probability distributions. For each source-language/target-language pair L_1 and L_2, we obtain the similarity of each of the 114 documents in L_1 to each of the 114 documents in L_2. We found that similarity here is best computed using the Jensen-Shannon divergence

	Tokenization	
Weighting Scheme	Word	Morpheme
Unweighted	0.505	0.544
$\log p(w\|L)$	0.616	0.641
PMI	0.612	**0.686**

Table 1: Summary of comparison results. This table shows the average precision at one document (P1) for each of the tokenization and weighting schemes we evaluated. Detailed results are presented in Table 2.

(Lin, 2002) and so this measure was used in all tests. Ultimately, the measure of how well a particular method performs is average precision at 1 document (P1). Among the various measurements for evaluating the performance of IR systems (Salton and McGill (1983), van Rijsbergen (1979)), this is a fairly standard measure. For a particular source-target pair, this is the percentage (out of 114 cases) where a document in L_1 is most similar to its mate in L_2. With 5 languages, there are 25 source-target pairs, and we can also calculate average P1 across all language pairs. Here, we average across 114 $\times$ 25 (or 2,850) cases. This is why even small differences in P1 can be statistically significant.

5 Results

First, we present a summary of our results in Table 1 which clearly demonstrates that it is better in LDA to use some kind of weighting scheme rather than the uniform weights in the standard LDA formulation from Eq. 1. This is true whether tokenization is by word or by morpheme. All increases from the baseline precision at 1 document (0.505 and 0.544 respectively), whether under log or PMI weighting, are highly significant ($p < 10^{-11}$). Furthermore, all increases in precision when moving from word-based to morphology-based tokenization are also highly significant ($p < 5 \times 10^{-5}$ without weighting, $p < 5 \times 10^{-3}$ with log-weighting, and $p < 2 \times 10^{-15}$ with PMI weighting). The best result overall, where P1 is 0.686, is obtained with morphological tokenization and PMI weighting (parallel to the results in (Chew et al., 2010) with LSI), and again the difference between this result and its nearest competitor of 0.641 is highly significant ($p < 3 \times 10^{-6}$). We return to comment below on lack of an increase in P1 when moving from log-weighting to PMI-weighting under word-based tokenization.

These results can also be broken out by language pair, as shown in Table 2. Here, it is apparent that Arabic, and to a lesser extent Russian, are harder languages in the IR problem at hand. Our intuition is that this is connected with the fact that these two languages have a more complex morphological structure: words are formed by a process of agglutination. A consequence of this is that single Arabic and Russian tokens can less frequently be mapped to single tokens in other languages, which appears to "confuse" LDA (and also, as we have found, LSI). The complex morphology of Russian and Arabic is also reflected in the type-token ratios for each language: in our English Bible, there are 12,335 types (unique words) and 789,744 tokens, a type-token ratio of 0.0156. The ratios for French, Spanish, Russian and Arabic are 0.0251, 0.0404, 0.0843 and 0.1256 respectively. Though the differences may not be explicable in purely statistical terms (there may be linguistic factors at play which cannot be reduced to statistics), it seems plausible that choosing a suboptimal term-weighting scheme could exacerbate any intrinsic problems of statistical imbalance. Considering this, it is interesting to note that the greatest gains, when moving from unweighted LDA to either form of weighted LDA, are often to be found where Russian and/or Arabic are involved. This, to us, shows the value of using a multilingual dataset as a testbed for our different formulations of LDA: it allows problems which may not be apparent when working with a monolingual dataset to come more easily to light.

We have mentioned that the best results are with PMI and morphological tokenization, and also that there is an increase in precision for many language of the pairs when morphological (as opposed to word-based) tokenization is employed. To us, the results leave little doubt that both weighting and morphological tokenization are independently beneficial. It appears, though, that morphology and weighting are also complementary and synergistic strategies for improving the results of LDA: for example, a suboptimal approach in tokenization may at best place an upper bound on the overall precision achievable, and perhaps at worst undo the benefits of a good weighting scheme. This may explain the one apparently anomalous result, which is the lack of an increase in

	Original Words					Morphological Tokenization					
	EN	ES	RU	AR	FR	EN	ES	RU	AR	FR	
EN	1.000	0.500	0.447	0.132	0.816	1.000	0.500	0.658	0.211	0.640	EN
ES	0.649	1.000	0.307	0.175	0.781	0.605	1.000	0.482	0.175	0.737	ES
LDA RU	0.430	0.316	1.000	0.149	0.430	0.553	0.421	1.000	0.272	0.553	RU
AR	0.070	0.149	0.114	1.000	0.096	0.123	0.105	0.228	1.000	0.114	AR
FR	0.781	0.693	0.421	0.175	1.000	0.693	0.640	0.667	0.211	1.000	FR
EN	1.000	0.518	0.518	0.228	0.658	1.000	0.675	0.561	0.219	0.754	EN
ES	0.558	1.000	0.605	0.254	0.763	0.711	1.000	0.570	0.289	0.860	ES
Log-WLDA RU	0.605	0.615	1.000	0.298	0.702	0.684	0.667	1.000	0.289	0.728	RU
AR	0.404	0.430	0.526	1.000	0.439	0.430	0.439	0.535	1.000	0.404	AR
FR	0.667	0.667	0.658	0.281	1.000	0.711	0.667	0.561	0.289	1.000	FR
EN	1.000	0.579	0.658	0.272	0.702	1.000	0.719	0.658	0.342	0.851	EN
ES	0.596	1.000	0.623	0.246	0.693	0.816	1.000	0.675	0.272	0.798	ES
PMI-WLDA RU	0.649	0.579	1.000	0.307	0.693	0.702	0.693	1.000	0.360	0.772	RU
AR	0.351	0.368	0.421	1.000	0.351	0.456	0.474	0.509	1.000	0.377	AR
FR	0.693	0.667	0.605	0.254	1.000	0.825	0.772	0.719	0.333	1.000	FR

Table 2: Full results for precision at one document for all combinations of LDA, Log-WLDA, PMI-WLDA, word tokenization and morphological tokenization.

precision moving from log-WLDA to PMI-WLDA under word-based tokenization: if word-based tokenization is suboptimal, PMI weighting cannot compensate for that. Effectively, for best results, the right strategies have to be pursued with respect *both* to morphology *and* to weighting.

Finally, we can illustrate the differences between weighted and unweighted LDA in another way. As discussed earlier, each topic in LDA is a probability distribution over terms. For each topic, we can list the most probable terms in decreasing order of probability; this gives a sense of what each topic is 'about' and whether the groupings of terms appear reasonable. Since we use 200 topics, an exhaustive listing is impractical here, but in Table 3 we present some representative examples from unweighted LDA and PMI-WLDA that we judged to be of interest. It appears to us that the groupings are not perfect under either LDA or PMI-WLDA; under both methods, we find examples of rather heterogeneous topics, whereas we would like each topic to be semantically focused. Still, a comparison of the output with LDA and PMI-WLDA sheds some light on why PMI-WLDA makes it less necessary to remove stopwords. Note that all words listed for the top two topics under LDA would commonly be considered stopwords. This might also be true of the words in topic 1 for PMI-WLDA, but in the latter case, the topic is actually one of the most semantically focused in that the top words have a clear semantic connection to one another. This cannot be said of topics 1 and 2 in LDA. For one thing, many of the same terms that appear in topic 1 reappear in topic 2, making the two topics hard to distinguish from one another. Secondly, the terms have only a loose semantic connection to one another: 'the', 'and', and 'of' are all high-frequency and likely to co-occur, but they are different parts of speech and have very different functions in English. One might say that topics 1 and 2 in LDA are a rag-bag of high-frequency words, and it is unsurprising that these topics do little to help characterize documents in our cross-language IR task. The same cannot be said of any of the top 5 topics in PMI-WLDA. We believe this illustrates well, and at a fundamental level, why weighted forms of LDA work better in practice than unweighted LDA.

6 Conclusion

We have conducted a series of experiments to evaluate the effect of different weighting schemes on Latent Dirichlet Allocation. Our results demonstrate, perhaps contrary to the conventional wisdom that weighting is unnecessary in LDA, that weighting schemes (and other pre-processing strategies) simi-

	Weighting Scheme									
	LDA (no weighting)					PMI-WLDA				
Topic	1	2	3	4	5	1	2	3	4	5
Terms	the	the	vanité	as	cárcel	under	city	coeur	sat	colère
	et	de	vanidad	comme	prison	sous	ville	heart	assis	ira
	and	et	vanity	como	السجن	под	ciudad	corazón	vent	wrath
	los	of	باطل	как	prison	тот	лабиринта	сердце	wind	anger
	и	and	суета	un	темницу	debajo	город	сердца	viento	furor
	y	y	aflicción	a	prisonniers	ombre	twelve	قلبه	sentado	гнев
	les	de	poursuite	one	темницы	bases	douze	قلب	ветер	fureur
	á	и	الباطل	لشما	bound	basas	doce	قلبي	الريح	غضب
	de	la	prédicateur	une	prisión	sombra	دينة	قلبك	sitting	гнева
	of	la	وقبض	واحد	prisoners	dessous	города	сердцем	сел	contre

Table 3: Top 10 terms within top 5 topics for each of LDA and PMI-WLDA. Terms that appear twice within the same topic (e.g. 'la' in LDA topic 2) are words from different languages with the same spelling (here Spanish and French).

lar to those commonly employed in other approaches to IR (such as LSI) can significantly improve the performance of a system. Our approach also runs counter to the standard position in LDA that it is necessary or desirable to remove stopwords as a preprocessing step, and we have presented an alternative approach of applying an appropriate weighting scheme within LDA. This approach is preferable because it is considerably less ad-hoc than the construction of stoplists. We have shown mathematically how alternative weighting schemes can be incorporated into the Gibbs sampling model. We have also demonstrated that, far from being arbitrary, the introduction of weighting into the LDA model has a solid and rational basis in information and probability theory, just as the basic LDA model itself has.

In future work, we would like to explore further enhancements to weighting in LDA. There are many variants which can be considered: one example is the incorporation of word order and context through an n-gram model based on conditional probabilities. We also aim to evaluate LDA against LSI with a view to establishing whether one can be said to outperform the other consistently in terms of precision, with appropriate settings held constant. Finally, we would like to determine whether other techniques which have been shown to benefit LSI can also be usefully brought to bear in LDA, just as we have shown here in the case of term weighting.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, pages 993–1022.

Sourav Chatterji and Lior Pachter. 2004. Multiple Organism Gene Finding by Collapsed Gibbs Sampling. In *RECOMB '04: Proceedings of the eighth annual international conference on Research in computational molecular biology*, pages 187–193, New York, NY, USA. ACM.

Peter A. Chew and Ahmed Abdelali. 2007. Benefits of the 'Massively Parallel Rosetta Stone': Cross-Language Information Retrieval with Over 30 Languages. In Association for Computational Linguistics, editor, *Proceedings of the 45th meeting of the Association of Computational Linguistics*, pages 872–879.

Peter A. Chew, Brett W. Bader, Stephen Helmreich, Ahmed Abdelali, and Stephen J. Verzi. 2010. An Information-Theoretic, Vector-Space-Model Approach to Cross-Language Information Retrieval. *Journal of Natural Language Engineering*. Forthcoming.

Philipp Cimiano, Antje Schultz, Sergej Sizov, Philipp Sorg, and Steffen Staab. 2009. Explicit Versus Latent Concept Models for Cross-Language Information Retrieval. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1513–1518.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Susan T. Dumais. 1991. Improving the Retrieval of Information from External Sources. *Behavior Research Methods, Instruments and Computers*, 23(2):229–236.

Stuart Geman, Donald Geman, K. Abend, T. J. Harley, and L. N. Kanal. 1993. Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images*. *Journal of Applied Statistics*, 20(5):25–62.

J. Goldsmith. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2):153–198.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding Scientific Topics. In *Proceedings of the National Academy of Sciences USA*, volume 101, pages 5228–5235.

Thomas Hofmann. 1999. Probablistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual International SIGIR Conference*, pages 53–57.

Solomon Kullback and Richard A. Leibler. 1951. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22:49–86.

J. Lin. 2002. Divergence Measures based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, August.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining Multilingual Topics from Wikipedia. In *18th International World Wide Web Conference*, pages 1155–1155, April.

Daniel Ramage, Paul Heymann, Christopher D. Manning, and Hector Garcia-Molina. 2008. Clustering the Tagged Web. In *Second ACM International Conference on Web Search and Data Mining (WSDM 2009)*, November.

G. Salton and M. McGill, editors. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Biola University. 2006. The Unbound Bible. http://www.unboundbible.com.

C.J. van Rijsbergen. 1979. *Information Retrieval*. Butterworth-Heinemann.