Cross-Language Linking of News Stories on the Web Using Interlingual Topic Modelling

Wim De Smet
Department of Computer Science
K.U.Leuven
Celestijnenlaan 200A
Leuven, Belgium
wim.desmet@cs.kuleuven.be

Marie-Francine Moens
Department of Computer Science
K.U.Leuven
Celestijnenlaan 200A
Leuven, Belgium
marie-francine.moens@cs.kuleuven.be

Abstract

We have studied the problem of linking event information across different languages without the use of translation systems or dictionaries. The linking is based on interlingua information obtained through probabilistic topic models trained on comparable corpora written in two languages (in our case English and Dutch). To achieve this goal, we expand the Latent Dirichlet Allocation model to process documents in two languages. We demonstrate the validity of the learned interlingual topics in a document clustering task, where the evaluation is performed on Google News.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Stochastic Processes; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Machine translation*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*

General Terms

Algorithms, Measurement

Keywords

Latent Dirichlet Allocation, Event Detection

1. INTRODUCTION

Finding related documents across different languages is a valuable information retrieval task. For many languages we might not have access to translation dictionaries or a full translation system, or their use might be computationally too expensive to apply in an online search. In such situations it is useful to make use of only a limited translation of the documents, for instance by finding correspondences with regard to the topics discussed. In this case, it would be interesting to automatically learn topics cross-lingually from parallel or comparable corpora, the latter being available for many language pairs. In a parallel bilingual corpus, each document has an exact translation in the other language. A comparable corpus consists of pairs of documents that are not exact translations of each other, but that contain similar content.

In one example of an information retrieval task, namely processing news stories, it is often relevant to determine whether two stories report on the same event. An event is defined here as a well-specified happening at a certain moment in time (a single day or a short period) which deals with a certain set of topics (e.g., a hurricane and inundations, an earthquake and lack of drinking water) and involving some named entities. Those entities are, for instance, the actors (such as the names of the leading persons or companies) and the location where the event occurred.

In this article we focus on the clustering of Google News stories coming from different language sources (we use the words "story" and "document" interchangeably). Probabilistic topic models such as Latent Dirichlet Allocation (LDA) [5] allow for describing a document in terms of a probabilistic distribution over topics, and these topic in terms of distributions over words. The original LDA setting creates word distributions in a single language. An initial idea is to train two topic models separately, one for each language in the corpus. However, an important property of LDA is *exchangeability*, which dictates that the ordering of topics in a topic model can be permutated without affecting the validity of the model. Therefore, when learning two models independently, we cannot guarantee that the topic representations will be comparable. As a solution, we extended the algorithm to learn two sets of topics, each for a different language, simultaneously. This way, we can create topic distributions for documents in both languages that can be compared.

In our experiments we consider two languages, i.e. English and Dutch. We train the word distributions of the bilingual topics based on a bilingual corpus. Next, we show that using the bilingual topics and the extracted named entities, we can cluster documents on the same event across the two languages without the use of any translation system or dictionary.

The contributions of the paper are the expansion of the Latent Dirichlet Allocation model with regard to the computation of interlingual topic models, and the cross-language clustering of events using the interlingual topics together with entity names.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 describes our methodology. We report on results, their evaluation and discussion

SWSM'09, November 6, 2009, Hong Kong, China.

in section 4. In section 5, we present our conclusions and aims for future research.

2. RELATED RESEARCH

Probabilistic topic models are already around for some years. The first model was probabilistic Latent Semantic Analysis (pLSA), developed by Hofmann [10]. Given the word distribution of a document collection, the topic distribution of each document and the word distribution for each topic are computed using approximate inference methods, whereby a topic is seen as a hidden variable and the number of topics is a priori defined. Because of the inability of pLSA to infer the topic distribution of a new document that is not part of the training corpus (apart from some limited folding in), and a number of parameter estimations that rises linearly with the number of documents in the training corpus, the Latent Dirichlet Allocation model (LDA) proposed by [5] is now one of the most popular methods for inferring latent topics.

Interlingual topic models were developed based on algebraic models considering Latent Semantic Analysis (LSA) [9]. Based on singular value decomposition of the term by document matrix of a document collection followed by dimensionality reduction, in a LSA model documents are represented in a lower dimensional vector space, where each vector component represents a topic. In the case of a parallel corpus (i.e., a document collection where each document has an exact translation in the other language), each document is concatenated with its counterpart document in the other language to form an interlingua term by document matrix, from which interlingual topic components can form the lower dimensional representation of the documents. Such representations were used in a cross-language retrieval setting [4, 15, 6, 18, 7] and document clustering [17]. A method based on LSA, but only using a short set of manually gathered comparable documents was presented in [23].

The idea of interlingual probabilistic topic models trained on comparable corpora is quite new. [26] use bilingual topic models trained on parallel corpora for word alignment and machine translation, where the bilingual topic models better capture the context in which a word is translated. Recently, a approach similar to ours was presented in [19].

Event detection has received a substantial interest in information retrieval research, often as part of topic detection and tracking (TDT) tasks. Early work on retrospective event detection based on a hierarchical agglomerative clustering (group average clustering) is done by [24] (building further on [8]). The events are clustered based on lexical (single words) similarity of the documents and temporal proximity. The temporal proximity parameter avoids clustering documents that are too far apart in time. Many different studies on event detection followed these initial initiatives (see [1] for the main approaches). Many of them rely on a vector space representation of the documents, where more recent approaches make a distinction between named entities and non named entity words (e.g., [12]). In such a scheme each term type might receive a different weight, possibly learned from a training corpus [25].

Probabilistic models for representing events in documents are scarce. [2] use a simple probabilistic language model as a document representation. [14] build a probabilistic generative model for retrospective news events detection, where an event generates persons, locations, keywords as named entities apart from a time pointer. Other research on integrating named entities in an event detection task include [16], [25], where [25] demonstrated correlations between named entity types and news classes. [21] demonstrated the value of splitting the similarity metric used for event clustering into two separate components respectively based on the similarity between topics and the similarity between named entities. It is along these lines that we want to perform cross-language event detection.

Cross-language event detection is a novel research domain, which recently received some attention in the homeland security realm. In this article we propose interlingual topic modeling for cross-language story linking according to event. [13] report on multilingual topic tracking, but still rely on standard machine translation. Cross-lingual news topic tracking was reported based on cognates (words that are spelled identically over different languages), named entities spelled identically and supervised classification with interlingual codes [20]. The method that we present is completely unsupervised.

3. METHODOLOGY

In this section, we first present a short description of the original LDA algorithm. Then, we adapt it to our bilingual needs. Finally, we present the setting in which we apply our adapted LDA-model, namely bilingual event detection.

3.1 LDA for monolingual documents

Latent Dirichlet Allocation, as described in [5], is a generative process that creates a set of documents. First, a corpus is associated with two variables associated: α and β. α is the k-dimensional parameter of a Dirichlet distribution from which, for each document, we sample a mixture over k topics, called θ. Then, to each of the N word positions in the document a topic z_n is assigned by sampling from θ. When a word position's topic is known, the word w_n itself is selected according to $p(w_n|\beta, z_n)$, where β defines for each z_n a multinomial distribution over the vocabulary. In summary:

1. Choose $\theta \sim Dir(\alpha)$.
2. For each of the $N_{\mathbf{t}}$ word positions t_n:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$.
   (b) Choose a word t_n from $p(t_n|z_n, \beta)$, a multinomial probability. conditioned on the topic z_n.

The sampling of $N_{\mathbf{t}}$ is usually left out of the equation.

3.2 LDA for multilingual documents

The expanded model that computes interlingual topic distributions in two languages follows the original model of [5] very closely. Figure 1 shows a graphic representation.

1. Choose $\theta \sim Dir(\alpha)$.
2. For each of the $N_{\mathbf{t}}$ word positions t_n:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$.
   (b) Choose a word t_n from $p(t_n|z_n, \beta)$, a multinomial probability. conditioned on the topic z_n.
3. For each of the $N_{\mathbf{v}}$ word positions v_n:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$.

(b) Choose a word v_n from $p(v_n|z_n, \nu)$, a multinomial probability. conditioned on the topic z_n.

Every document has two different kinds of features: t_n and v_n. t_n are words in one language and v_n are words in the other language. t_n are sampled according to a multinomial distribution conditioned on z_n and the global variable β. In the rest of the paper, $p(t_n|z_n, \beta)$ is taken to be β_{ij}, if t_n has the vocabulary (V_t) index j and $z_n = i$. For the words in the other language we follow the same reasoning, so that $p(v_n|z_n, \nu)$ is taken to be ν_{ij}, if v_n has the vocabulary (V_v) index j and $z_n = i$.

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of $N_\mathbf{t} + N_\mathbf{v}$ topics $\mathbf{z}$, and a set of $N_\mathbf{t}$ text words $\mathbf{t}$ and $N_\mathbf{v}$ words $\mathbf{v}$ in the other language is given by:

$$p(\theta, \mathbf{z}, \mathbf{t}, \mathbf{v}|\alpha, \beta, \nu) = p(\theta|\alpha) \prod_{n=1}^{N_\mathbf{t}} p(z_n|\theta)p(t_n|z_n, \beta) \prod_{n=1}^{N_\mathbf{v}} p(z_n|\theta)p(v_n|z_n, \nu).$$

Integrating over θ and summing over $\mathbf{z}$, we obtain the marginal distribution of a document:

$$p(\mathbf{t}, \mathbf{v}|\alpha, \beta, \nu) = \int p(\theta|\alpha)(\prod_{n=1}^{N_\mathbf{t}} \sum_{z_n} p(z_n|\theta)p(t_n|z_n, \beta) \prod_{n=1}^{N_\mathbf{v}} \sum_{z_n} p(z_n|\theta)p(v_n|z_n, \nu)d\theta.$$

Training an LDA model means finding the values for parameters β, ν and θ that maximize the product of the probability over all documents. We obtain this by expanding the variational inference method as implemented by Blei[1]. In the appendix, we give the full derivations.

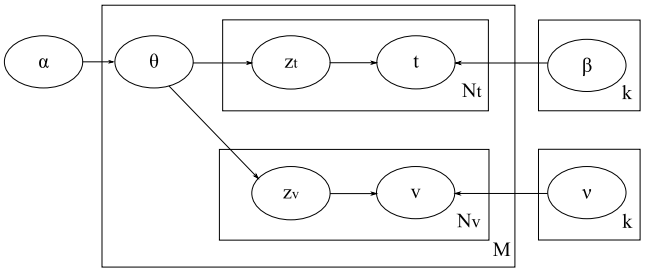

Figure 1: Plate model of the interlingual topic model.

We train from a comparable corpus, which contains pairs of documents in both languages. Each pair of documents discusses the same topics. We can also use a parallel corpus for constructing the interlingual topic model, but parallel corpora are less frequently available.

3.3 Test case: Event detection

We want to test our interlingual probabilistic topic model for clustering news stories written in English and Dutch into groups of stories that describe the same event. If we succeed, we have defined a method for automatically linking event stories across languages. An event can be seen as a mixture of topics, where some topics are prominently and others only marginally present. We follow here the method for event detection, described in [21], but now applied with an interlingual topic model.

When LDA is trained on the documents' full texts, the named entities (e.g., person, location or organization names) are part of the topic distributions. This has the undesirable property that entities that were not apparent in the training set (which, given the dynamic nature of news, occurs often) can not influence the topic inference of a new event. Therefore, we train the LDA model on documents where the entities, detected with a simple recognizer which relies on capitalization patterns in text, have been removed first. To represent the named entities of a document, we consider the named entities actually present in the document and estimate the probability of a named entity by smoothed maximum likelihood estimation in the document.

Documents (in our case Google News stories) are now represented by a probability distribution over topics, and a probability distribution over entities. If we want to cluster the documents according to event, we need a dissimilarity function. For both representatoins, We use the symmetric Kullback-Leibler divergence of the n-dimensional probability distributions d_i and d_j, defined as:

$$KL(d_i, d_j) = \frac{1}{2}\Big(\sum_{l=1}^{n} d_i^l \log(\frac{d_i^l}{d_j^l}) + \sum_{l=1}^{n} d_j^l \log(\frac{d_j^l}{d_i^l})\Big)$$

For entities, d_i is the smoothed term vector normalized by its sum, for LDA generated topics it is the distribution associated with the document. n can be either the number of topics or number of entities.

In order to obtain a final dissimilarity function, the dissimilarities in topic distribution and entity distribution are combined by the maximum function, which proved to yield the best results in monolingual event detection [21]:

$$dis(d_i, d_j) = \max_k dis(A_{d_i}^k, A_{d_j}^k), k = 1 \rightarrow N$$

where N is the number of content representations or *aspects* the document is split into, which in our case equals 2 (topic distribution and entity distribution). A_d^k is the kth content representation of d.

The *max*-function ensures that two documents are dissimilar when at least one of the aspects has dissimilar distributions: if two documents differ too much in one aspect, then it is irrelevant whether the other aspects are close or not. In an event setting, this translates into the following: if we detect different actors or locations, then we assume that we deal with different events, even when their topics are similar. Analogically, events with different topics that happen at the same location will be treated as different events.

3.4 Clustering

The document dissimilarity $dis(d_i, d_j)$, which is a fused dissimilarity of topic and named entity dissimilarity, is used in a clustering algorithm. We used a hierarchical agglomerative clustering with complete linkage, as it is mentioned in the literature as one of the best performing document clustering algorithms [22]. The hierarchical clustering algorithm does not require the number of clusters to be chosen a priori,

[1]http://www.cs.princeton.edu/ blei/lda-c/

	English	Dutch	Interlingual
Event detection relying on topic distribution	91.2% (17)	59.5% (12)	**63.2**% (23)
Event detection relying on entity distribution	80.1% (20)	**87.7**% (23)	56.7% (37)
Event detection relying on both topic and entity distribution	**94.1**% (23)	85.3% (23)	56.9% (48)
Baseline	94.5% (19)	100.0% (18)	56.5% (36)

Table 1: Results in terms of F_1 measure (B-Cubed) when using the topic distribution, entity distribution for monolingual and cross-lingual event detection. Number of found clusters is given in parentheses.

a very important property in our dynamic environment. We can use a fitness-condition on the clustering to create a natural, unsupervised stopping criterion. This natural clustering is the most logical extension of our unsupervised approach: the data provides the number of clusters itself.

For every document d_i in our corpus, we calculate its fitness in cluster C_i as the normalized difference between the distance of d_i to the second best cluster C_j, and the average distance of d_i to the other documents in C_i:

$$f(d_i) = \frac{b(d_i) - a(d_i)}{\max\{a(d_i), b(d_i)\}}$$

$$\text{where} \quad a(d_i) = \frac{1}{|C_i| - 1} \sum_{d_j \in C_i} dis(d_i, d_j)$$

$$\text{and} \quad b(d_i) = \arg\min_{C_j} \frac{1}{|C_j|} \sum_{d_j \in C_j} dis(d_i, d_j)$$

If C_i is a singleton cluster (containing only d_i), we assign $f(d_i)$ the default value 0. We search for the clustering that maximizes the average of f over all documents, over all possible stops in the hierarchy.

4. RESULTS

We will first give details on the datasets used in the evaluation of the event clustering. Then follows a short section on our clustering algorithms and cluster evaluation techniques. After that, we present results and their discussion.

4.1 Datasets

Training corpus.

As training set we used 7612 Wikipedia articles, selected via the "Random page" function of Wikipedia in its Dutch version, and then using the linked English counterpart documents. The Dutch texts are usually short, containing 84 words on average with a standard variation of 157 words. The English texts contain on average 968 words and standard deviation of 1443 words. Both corpora contain outliers in length. We assumed 100 LDA topics to be present in this data set.

Test corpus.

As test corpus we randomly selected 18 recent events in Google news in the period of July 16-18 2009, forming 18 clusters of English and Dutch news documents. The 18 clusters contain in total 50 documents written in English, with an average of 347.0 words and standard deviation of 254.6 words, and 60 documents written in Dutch, with an average of 72.5 words and standard deviation of 43.2 words.

4.2 Evaluation metrics

The evaluation of our clustering is done using the B-Cubed metric [3]. Let C_i be the symbol for the cluster that document d_i gets clustered in, and M_i be its manual cluster (i.e. from the ground truth). The B-Cubed metric then calculates for each document its precision (how many of the other documents in its automatic cluster should be in it?) as $\frac{|C_i \cap M_i|}{|C_i|}$, and its recall (how many of the documents in its manual cluster are in its automatic cluster?) as $\frac{|C_i \cap M_i|}{|M_i|}$. The total clustering precision and recall are taken as the average over all documents.

Our main remark on the B-Cubed metric is the fact that it rewards a singleton clustering (each document in its own cluster) with a precision of 100%, as no document is clustered together with an unrelated one. Of course, recall will be very low in that case. Therefore we present the F1 values, as these give a clear view on both precision and recall.

4.3 Results of the interlingual topic construction

The training of the interlingual topic model on the English-Dutch bilingual comparable corpus gave intuitively very good results: a visual inspection of the words from both languages in each topic showed a strong semantic relation between the bilingual topics, with many translations appearing within the top 100 words (see example in figure 2).

4.4 Results of cross-lingual event clustering

We report in table 1 on a clustering of the documents according to their event. We cluster the documents considering the found topic distributions, the entity distributions and both distributions combined. We cluster in a monolingual setting, i.e., considering each language, in our case English and Dutch, and in a cross-lingual setting, i.e., clustering the English and Dutch documents according to event solely relying on the interlingual topics and entity names, which might be spelled differently in the two languages.

To compare, we also provide baseline results. In contrast to our unsupervised algorithm, these results are obtained by using a translation dictionary to expand the words in one language to several possible translations in the other language (disregarding context). The cosine metric is then used to measure the distance between the documents. We use these distances in the same clustering algorithm. The reported performance is the average of using the English and translated Dutch documents, and vice versa. The same metric and clustering is also applied to the monolingual datasets.

In the monolingual setting the event clustering is quite accurate. In English the event clustering yields best results based on the topic models and in Dutch the entity distributions most accurately define the clusters. Combining topic and entity distributions yield very good event de-

auto (car)	car	literatuur (literature)	literature	gebouw (building)	building
modellen (models)	engine	eeuw (century)	poetry	meter (meter)	cort
model (model)	model	god (god)	works	eeuw (century)	buildings
rpm (rpm)	cars	man (man)	literary	bisschop (bishop)	built
motor (engine)	kw	verhaal (story)	goliath	kasteel (castle)	garden
productie (production)	production	werk (work)	dutch	gebouwd (built)	museum
nieuwe (new)	models	den (/)	poets	stad (city)	palace
gebouwd (built)	cc	teksten (texts)	book	museum (museum)	construction
verkocht (sold)	door	bekend (famous)	poems	theater (theatre)	tower
motoren (engines)	rear	grote (great)	period	tuin (garden)	theatre

Figure 2: 10 most probable words in Dutch and English. For the Dutch words, we have given the English translation. It is clear that in both languages words are chosen that are related to respectively *cars*, *literature* and *architecture*.

tection (94.1% and 85.3 F_1 measure for respectively English and Dutch). It should be noted that we show here the results for the cluster with the highest fitness., i.e., the algorithm itself decides the number of clusters. The monolingual baseline method performs slightly better in the English case, and a lot better in the Dutch case. This is caused by little variation in the wording used for the Dutch news articles.

As expected, the cross-lingual setting is more difficult without relying on any translation system or dictionary. Still, the use of the interlingual topic models is attaining 63.2% F_1 measure for cross-language applications without relying on machine translation systems or dictionaries. Clustering based on entities is much lower than in the monolingual case, as many names are translated (geographical names for example), or spelled differently in the two languages. Whereas the baseline method performed better for a monolingual case, the topic model outperforms the baseline method, when ignoring the noisy named entities. A method that manages to translate these entities will likely improve the performance of our method further.

When performing a limited error analysis, we found that the Dutch texts are generally shorter containing less content, which might disturb the cross-lingual topic inference as their topic distribution is less outspoken with regard to the main topics.

5. CONCLUSION

In this paper we have presented a model for interlingual probabilistic topics by expanding the Latent Dirichlet Allocation model of [5]. We have successfully applied the model in the task of cross-language event detection in Google News. We demonstrated that the model contributes to the event clustering task without relying on a machine translation system or bilingual dictionaries, in a realistic situation where the model is trained on a bilingual comparable corpus.

In future work, we want to consider additional languages when training the interlingual probabilistic topic model and use the model in cross-language information retrieval. Also, translation of named entities should prove useful in disambiguating stories.

6. APPENDIX

This appendix is an extension of the appendix in [5], to include the extra parameters in our model.

The key inferential problem that we need to solve in order to use LDA is that of computing the posterior distribution

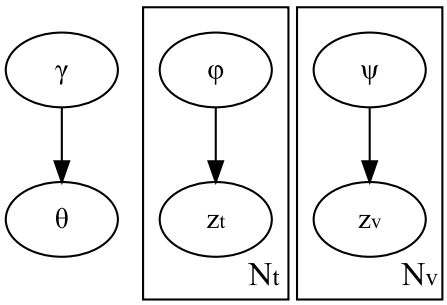

Figure 3: Plate model of the interlingual variational model

of the hidden variables given a document:

$$p(\theta, \mathbf{z}|\mathbf{t}, \mathbf{v}, \alpha, \beta,) = \frac{p(\theta, \mathbf{z}, \mathbf{t}, \mathbf{v}, \alpha, \beta, \nu)}{p(\mathbf{t}, \mathbf{v}|\alpha, \beta, \nu)}$$

Unfortunately, this distribution is intractable to compute in general. Indeed, to normalize the distribution we marginalize over the hidden variables and write the equation in terms of the model parameters

$$p(\mathbf{t}, \mathbf{v}|\alpha, \beta) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \int (\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}) (\prod_{n=1}^{N_\mathbf{t}} \sum_{i=1}^{k} \prod_{j=1}^{V_t} (\theta_i \beta_{ij})^{\mathbf{t}_n^j}) (\prod_{n=1}^{N_\mathbf{v}} \sum_{i=1}^{k} \prod_{j=1}^{V_v} (\theta_i \nu_{ij})^{\mathbf{v}_n^j}) d\theta,$$

a function which is intractable due to the coupling between θ and β in the summation over latent topics. To solve this problem, we use variational inference to approximate the distribution. The problematic coupling between θ and β arises due to the edges between θ, $\mathbf{z}$, $\mathbf{t}$ and $\mathbf{v}$. By decoupling these edges and introducing new nodes, we get a solution which is tractable. The graphical model is shown in figure 3. This variational distribution

$$q(\theta, \mathbf{z}|\gamma, \phi, \psi) = q(\theta|\gamma) \prod_{n=1}^{N_\mathbf{t}} q(z_n|\phi_n) \prod_{n=1}^{N_\mathbf{v}} q(z_n|\psi_n)$$

creates independency between $\mathbf{z}$ and θ.

We need to find the γ, ϕ and ψ which optimize the approximation of $p(\theta, \mathbf{z}, \mathbf{t}, \mathbf{v}|\alpha, \beta, \nu)$ by $q(\theta, \mathbf{z}|\gamma, \phi, \psi)$. Following [11], we begin by bounding the log likelihood of a document

using Jensens inequality. Omitting the parameters γ, ϕ and ψ for simplicity, we have:

$$\begin{aligned}
\log(\mathbf{t},\mathbf{v}|\alpha,\beta,\nu) &= \log\int\sum_{\mathbf{z}} p(\theta,\mathbf{z},\mathbf{t},\mathbf{v}|\alpha,\beta,\nu)d\theta \\
&= \log\int\sum_{\mathbf{z}} \frac{p(\theta,\mathbf{z},\mathbf{t},\mathbf{v}|\alpha,\beta,\nu)q(\theta,\mathbf{z})}{q(\theta,\mathbf{z})}d\theta \\
&\geq \int\sum_{\mathbf{z}} q(\theta,\mathbf{z})\log p(\theta,\mathbf{z},\mathbf{t},\mathbf{v}|\alpha,\beta,\nu) \\
&\quad -\int\sum_{\mathbf{z}} q(\theta,\mathbf{z})\log q(\theta,\mathbf{z}) \\
&= E_q[\log p(\theta,\mathbf{z},\mathbf{t},\mathbf{v}|\alpha,\beta,\nu)] - E_q[\log q(\theta,\mathbf{z})],
\end{aligned}$$

where E_q is the expected value according to the variational distribution. We denote the right hand side of as $L(\gamma,\phi,\psi;\alpha,\beta,\nu)$, the lower bound on the log likelihood of a document. As it is a lower bound, we need to maximalize it with respect to parameters γ, ϕ and ψ in order to have the best approximation of p. As $p(\theta,\mathbf{z},\mathbf{t},\mathbf{v}|\alpha,\beta,\nu)$ can be factorized into $p(\theta|\alpha)\cdot p(\mathbf{z}|\theta)\cdot p(\mathbf{t}|\mathbf{z},\beta)\cdot p(\mathbf{v}|\mathbf{z},\nu)$, then the log can be factorized as $\log p(\theta|\alpha) + \log p(\mathbf{z}|\theta) + \log p(\mathbf{t}|\mathbf{z},\beta) + \log p(\mathbf{v}|\mathbf{z},\nu)$. Similarly, $\log q(\theta,\mathbf{z})$ becomes $\log q(\theta)+\log q(\mathbf{z})$. Using the property that for any two stochastic variables $E(X,Y) = E(X) + E(Y)$, we can factorize the lower bound as:

$$\begin{aligned}
L(\gamma,\phi,\psi;\alpha,\beta,\nu) =& E_q[\log p(\theta|\alpha)] \\
&+ E_q[\log p(\mathbf{z}|\theta)] \\
&+ E_q[\log p(\mathbf{t}|\mathbf{z},\beta)] + E_q[\log p(\mathbf{v}|\mathbf{z},\nu)] \\
&+ E_q[\log q(\theta)] + Eq[\log q(\mathbf{z})]
\end{aligned} \tag{1}$$

In several occasions, we use the following formula:

$$E[\log\theta_i|\alpha] = \Psi(\alpha_i) - \Psi(\sum_{i=1}^{k}\alpha_j),$$

which translates for the variational distribution into

$$E_q[\log\theta_i|\gamma] = \Psi(\gamma_i) - \Psi(\sum_{i=1}^{k}\gamma_j)$$

We expand each of the terms in equation 1 according to the parameters α, β and ν of the model, and the variational parameters γ, ϕ and ψ:

$$\begin{aligned}
E_q[\log p(\theta|\alpha)] &= E_q\Big[\log\Big(\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}\big(\theta_i^{\alpha_i-1}\big)\Big)\Big] \\
&= E_q\Big[\log\Gamma(\sum_{i=1}^{k}\alpha_i) - \log\prod_{i=1}^{k}\Gamma(\alpha_i) + \log\prod_{i=1}^{k}\theta_i^{\alpha_i-1}\Big] \\
&= \log\Gamma(\sum_{i=1}^{k}\alpha_i) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) \\
&\quad + \sum_{i=1}^{k}(\alpha_i-1)\big(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\big)
\end{aligned}$$

$$\begin{aligned}
E_q[\log p(\mathbf{t}|\mathbf{z},\beta)] &= E_q\Big[\log\prod_{n=1}^{N_\mathbf{t}}\prod_{i=1}^{k}\prod_{j=1}^{V_t}\beta_{ij}^{(z_n^i\cdot t_n^j)}\Big] \\
&= E_q\Big[\sum_{n=1}^{N_\mathbf{t}}\sum_{i=1}^{k}\sum_{j=1}^{V_t} z_n^i\cdot t_n^j\cdot\log\beta_{ij}\Big] \\
&= \sum_{n=1}^{N_\mathbf{t}}\sum_{i=1}^{k}\sum_{j=1}^{V_t}\phi_{ni}\cdot t_n^j\cdot\log\beta_{ij}
\end{aligned}$$

$$\begin{aligned}
E_q[\log p(\mathbf{v}|\mathbf{z},\nu)] &= E_q\Big[\log\prod_{n=1}^{N_\mathbf{v}}\prod_{i=1}^{k}\prod_{j=1}^{V_v}\nu_{ij}^{(z_n^i\cdot v_n^j)}\Big] \\
&= E_q\Big[\sum_{n=1}^{N_\mathbf{t}}\sum_{i=1}^{k}\sum_{j=1}^{V_v} z_n^i\cdot v_n^j\cdot\log\nu_{ij}\Big] \\
&= \sum_{n=1}^{N_\mathbf{t}}\sum_{i=1}^{k}\sum_{j=1}^{V_v}\phi_{ni}\cdot v_n^j\cdot\log\nu_{ij}
\end{aligned}$$

$$\begin{aligned}
E_q[\log p(\mathbf{z}|\theta)] &= E_q\Big[\log\prod_{n=1}^{N_\mathbf{t}+N_\mathbf{v}}\prod_{i=1}^{k}\theta_i^{z_n^i}\Big] \\
&= E_q\Big[\sum_{n=1}^{N_\mathbf{t}+N_\mathbf{v}}\sum_{i=1}^{k}\log\theta_i^{z_n^i}\Big] \\
&= \sum_{n=1}^{N_\mathbf{t}+N_\mathbf{v}}\sum_{i=1}^{k}E_q[z_n^i\cdot\log\theta_i] \\
&= \sum_{n=1}^{N_\mathbf{t}}\sum_{i=1}^{k}\phi_{ni}\Big(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\Big) \\
&\quad + \sum_{n=1}^{N_\mathbf{v}}\sum_{i=1}^{k}\psi_{ni}\Big(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\Big)
\end{aligned}$$

$$\begin{aligned}
E_q[\log q(\theta)] &= E_q\Big[\log\frac{\Gamma(\sum_{i=1}^{k}\gamma_i)}{\prod_{i=1}^{k}\Gamma(\gamma_i)}\prod_{i=1}^{k}\theta_i^{(\gamma_i-1)}\Big] \\
&= E_q\Big[\log\Gamma(\sum_{i=1}^{k}\gamma_i) - \log\prod_{i=1}^{k}\Gamma(\gamma_i) + \log\prod_{i=1}^{k}\theta_i^{(\gamma_i-1)}\Big] \\
&= \log\Gamma(\sum_{i=1}^{k}\gamma_i) - \sum_{i=1}^{k}\log\Gamma(\gamma_i) \\
&\quad + \sum_{i=1}^{k}(\gamma_i-1)\Big(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k}\gamma_j)\Big)
\end{aligned}$$

$$E_q[\log q(\mathbf{z})] = E_q\Big[\log\Big(\prod_{n=1}^{N_\mathbf{t}}\prod_{i=1}^{k}q(z_n|\phi_n)\prod_{n=1}^{N_\mathbf{v}}\prod_{i=1}^{k}q(z_n|\psi_n)\Big)\Big]$$

$$
\begin{aligned}
&= E_q\Big[\log\Big(\prod_{n=1}^{N_\mathbf{t}}\prod_{i=1}^{k}\phi_{ni}^{z_n^i}\prod_{n=1}^{N_\mathbf{v}}\prod_{i=1}^{k}\psi_{ni}^{z_n^i}\Big)\Big]\\
&= \sum_{n=1}^{N_\mathbf{t}}\sum_{i=1}^{k}E_q[z_n^i\cdot\log\phi_{ni}]+\sum_{n=1}^{N_\mathbf{v}}\sum_{i=1}^{k}E_q[z_n^i\log\psi_{ni}]\\
&= \sum_{n=1}^{N_\mathbf{t}}\sum_{i=1}^{k}\phi_{ni}\cdot\log\phi_{ni}+\sum_{n=1}^{N_\mathbf{v}}\sum_{i=1}^{k}\psi_{ni}\cdot\log\psi_{ni}
\end{aligned}
$$

Together:

$$
\begin{aligned}
L(\gamma,\phi,\psi;\alpha,\beta,\nu) =& \log\Gamma(\sum_{i=1}^{k}\alpha_i)-\sum_{i=1}^{k}\log\Gamma(\alpha_i)\\
&+\sum_{i=1}^{k}(\alpha_i-1)\big(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)\big)\\
&+\sum_{n=1}^{N_\mathbf{t}}\sum_{i=1}^{k}\phi_{ni}\Big(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)\Big)\\
&+\sum_{n=1}^{N_\mathbf{v}}\sum_{i=1}^{k}\psi_{ni}\Big(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)\Big)\\
&+\sum_{n=1}^{N_\mathbf{t}}\sum_{i=1}^{k}\sum_{j=1}^{V_t}\phi_{ni}\cdot t_n^j\cdot\log\beta_{ij}\\
&+\sum_{n=1}^{N_\mathbf{t}}\sum_{i=1}^{k}\sum_{j=1}^{V_v}\phi_{ni}\cdot v_n^j\cdot\log\nu_{ij}\\
&-\log\Gamma\big(\sum_{i=1}^{k}\gamma_i\big)+\sum_{i=1}^{k}\log\Gamma(\gamma_i)\\
&-\sum_{i=1}^{k}(\gamma_i-1)\Big(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)\Big)\\
&-\sum_{n=1}^{N_\mathbf{t}}\sum_{i=1}^{k}\phi_{ni}\cdot\log\phi_{ni}\\
&-\sum_{n=1}^{N_\mathbf{v}}\sum_{i=1}^{k}\psi_{ni}\cdot\log\psi_{ni}
\end{aligned}
$$

6.1 Variational multinomials

6.1.1 Multinomial for language t

To maximize the lower bound with respect to ϕ_{ni}, we take the terms in which ϕ_{ni} appears, and set the derivative to zero. The terms of $L(\gamma,\phi,\psi;\alpha,\beta,\nu)$ that only contain ϕ_{ni}, with a Lagrange constraint of $\sum_{i=1}^{k}\phi_{ni}=1$ become:

$$
\begin{aligned}
L_{[\phi_{ni}]} &= \phi_{ni}\big(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)\big)+\phi_{ni}\log\beta_{ix}-\phi_{ni}\log\phi_{ni}\\
&\quad+\lambda_n(\sum_{j=1}^{k}\phi_{ni}-1),
\end{aligned}
$$

where β_{ix} denotes the β for word x which appears at position n. Taking the derivative with respect to ϕ_{ni}, we obtain:

$$
\frac{\partial L_{[\phi_{ni}]}}{\partial\phi_{ni}}=\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)+\log\beta_{ix}-\log\phi_{ni}-1+\lambda
$$

Setting this derivative to zero yields the maximizing value of the variational parameter ϕ_{ni} :

$$
\phi_{ni}\propto\beta_{ix}\exp\big(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)\big)
$$

6.1.2 Multinomial for language v

Since ϕ_{ni} and ψ_{ni} do not occur in the same terms in the lower bound factorization, we can repeat the previous derivation, but substitute ϕ_{ni} with ψ_{ni} . This gives as the maximizing value for ψ_{ni}:

$$
\psi_{ni}\propto\nu_{ix}\exp\big(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)\big)
$$

6.2 Variational Dirichlet

The terms containing γ_i are:

$$
\begin{aligned}
L_{[\gamma_i]} &= \sum_{i=1}^{k}(\alpha_i-1)\big(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)\big)\\
&+ \sum_{n=1}^{N_\mathbf{t}}\phi_{ni}\big(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k})\big)+\sum_{n=1}^{N_\mathbf{v}}\psi_{ni}\big(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)\big)\\
&- \log\Gamma(\sum_{j=1}^{k}\gamma_j)+\log\Gamma(\gamma_i)-\sum_{i=1}^{k}(\gamma_i-1)\big(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)\big)
\end{aligned}
$$

This simplifies to:

$$
\begin{aligned}
L_{[\gamma_i]} =& \sum_{i=1}^{k}\big(\Psi(\gamma_i)-\Psi(\sum_{j=1}^{k}\gamma_j)\big)\cdot\\
&\big(\alpha_i+\sum_{n=1}^{N_\mathbf{t}}\phi_{ni}+\sum_{n=1}^{N_\mathbf{v}}\psi_{ni}\big)-\log\Gamma(\sum_{j=1}^{k}\gamma_j)+\log\Gamma(\gamma_i)
\end{aligned}
$$

We take the derivative with respect to γ_i :

$$
\begin{aligned}
\frac{\partial L_{[\gamma_i]}}{\partial\gamma_i} =& \Psi'(\gamma_i)(\alpha_i+\sum_{n=1}^{N_\mathbf{t}}\phi_{ni}+\sum_{n=1}^{N_\mathbf{v}}\psi_{ni}-\gamma_j)\\
&+\Psi'(\sum_{j=1}^{k})\gamma_i)\sum_{j=1}^{k}(\alpha_j+\sum_{n=1}^{N_\mathbf{t}}\phi_{nj}+\sum_{n=1}^{N_\mathbf{v}}\psi_{nj}-\gamma_j)
\end{aligned}
$$

Setting this equation to zero yields a maximum at:

$$
\gamma_i=\alpha_i+\sum_{n=1}^{N_\mathbf{t}}\phi_{ni}+\sum_{n=1}^{N_\mathbf{v}}\psi_{ni}
$$

6.3 Conditional multinomials

To maximize with respect to β, we again isolate its terms and add Lagrange multipliers:

$$
L_{[\beta]}=\sum_{d=1}^{M}\sum_{n=1}^{N_{dt}}\sum_{i=1}^{k}\sum_{j=1}^{V_t}\phi_{dni}\cdot v_{dn}^t\cdot\log\beta_{ij}+\sum_{i=1}^{k}\lambda_i\big(\sum_{j=1}^{V_t}\log\beta_{ij}-1\big).
$$

Taking the derivate with respect to β_{ij} and setting it to zero yields:

$$
\beta_{ij}\propto\sum_{d=1}^{M}\sum_{n=1}^{N_{dt}}\phi_{dni}\cdot t_{dn}^j
$$

Similarly, we find for ν_{ij} :

$$\nu_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_{dv}} \psi_{dni} \cdot v_{dn}^{j}$$

For the derivation with respect to α we point to the original LDA paper, as the terms there did not change.

7. REFERENCES

[1] J. Allan, V. Lavrenko, and R. Swan. *Explorations within Topic Tracking and Detection*, ir 20, pages 197–224. Kluwer Academic Publishers, 2002.

[2] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 314–321, New York, NY, USA, 2003. ACM.

[3] A. Bagga and B. Baldwin. Algorithms for scoring coreference chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566, 1998.

[4] M. W. Berry, S. T. Dumais, and G. W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Rev.*, 37(4):573–595, 1995.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[6] J. G. Carbonell, J. G. Yang, R. E. Frederking, R. D. Brown, Y. Geng, D. Lee, Y. Frederking, R. E, R. D. Geng, and Y. Yang. Translingual information retrieval: A comparative evaluation. *In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 708–714, 1997.

[7] P. A. Chew, B. W. Bader, T. G. Kolda, and A. Abdelali. Cross-language information retrieval using parafac2. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 143–152, New York, NY, USA, 2007. ACM.

[8] D. R. Cutting, J. O. Pedersen, D. Karger, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329, 1992.

[9] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[10] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of Uncertainty in Artificial Intelligence, UAI*, Stockholm, 1999.

[11] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical methods. In *Machine Learning*, pages 183–233. MIT Press, 1998.

[12] G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 297–304, New York, NY, USA, 2004. ACM.

[13] L. S. Larkey, F. Feng, M. Connell, and V. Lavrenko. Language-specific models in multilingual topic tracking. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 402–409, New York, NY, USA, 2004. ACM.

[14] Z. Li, B. Wang, M. Li, and W.-Y. Ma. A probabilistic model for retrospective news event detection. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 106–113, New York, NY, USA, 2005. ACM.

[15] M. Littman, S. T. Dumais, and T. K. Landauer. Automatic cross-language information retrieval using latent semantic indexing. In *Cross-Language Information Retrieval, chapter 5*, pages 51–62. Kluwer Academic Publishers, 1998.

[16] U. Makkonen, H. Ahonen-Myka, and Marko. Applying semantic classes in event detection and tracking. In *Proc. International Conference on Natural Language Processing (ICON'02)*, pages 175–183, 2002.

[17] B. Mathieu, R. Besançon, and C. Fluhr. Multilingual document clusters discovery. In *RIAO*, pages 116–125, 2004.

[18] T. Muramatsu and T. Mori. Integration of pLSA into probabilistic CLIR model. In *Proceedings of NTCIR-04*, 2004.

[19] X. Ni, J.-T. Sun, J. Hu, and Z. Chen. Mining multilingual topics from Wikipedia. In *18th International World Wide Web Conference*, pages 1155–1155, April 2009.

[20] B. Pouliquen, R. Steinberger, C. Ignat, and T. D. Groeve. Geographical information recognition and visualization in texts written in various languages. In *SAC*, pages 1051–1058, 2004.

[21] W. De Smet and M.-F. Moens. An aspect based document representation for event clustering. In *Proceedings of CLIN 19*.

[22] E. M. Voorhees. Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. Technical report, Ithaca, NY, USA, 1986.

[23] Y. Wu and D. W. Oard. Bilingual topic aspect classification with a few training examples. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 203–210, New York, NY, USA, 2008. ACM.

[24] Y. Yang, J. G. Carbonell, R. D. Brown, T. Pierce, B. T. Archibald, and X. Liu. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4):32–43, 1999.

[25] K. Zhang, J. Zi, and L. G. Wu. New event detection based on indexing-tree and named entity. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 215–222, New York, NY, USA, 2007. ACM.

[26] B. Zhao and E. P. Xing. Bitam: Bilingual topic admixture models for word alignment. In *ACL*, 2006.