Probabilistic Topic Modeling in Multilingual Settings: A Short Overview of Its Methodology and Applications

Ivan Vulić
KU Leuven
ivan.vulic@cs.kuleuven.be

Wim De Smet
KU Leuven
wdesmet@gmail.com

Jie Tang
Tsinghua University
jie.tang@tsinghua.cn.edu

Marie-Francine Moens
KU Leuven
marie-francine.moens@cs.kuleuven.be

Abstract

Probabilistic topic models are unsupervised generative models that model document content as a two-step generation process, i.e., documents are observed as mixtures of latent topics, while topics are probability distributions over vocabulary words. Recently, a significant research effort has been invested into transferring the probabilistic topic modeling concept from monolingual to multilingual settings. Novel topic models have been designed to work with parallel and comparable texts. We define the concept of multilingual probabilistic topic modeling and present a short high-level overview of the current research and methodology. As a representative example, we thoroughly describe a multilingual probabilistic topic model called bilingual LDA (BiLDA) trained on comparable data in the appendix. In the paper we provide a short overview of cross-lingual applications for which we utilized the model in our research so far.[1]

1 Introduction and Context

Probabilistic latent topic models such as probabilistic Latent Semantic Analysis (pLSA) [9] and Latent Dirichlet Allocation (LDA) [10] along with their numerous variants are well studied generative models for representing the content of documents in large document collections. They provide a robust and unsupervised framework for performing shallow latent semantic analysis of themes (or topics) discussed in text. The families of these latent topic models are all based upon the idea that there exist latent variables, i.e., *topics*, which determine how words in documents have been generated. Fitting such a generative model actually denotes finding the best set of those latent variables in order to explain the observed data. With respect to that generative process, documents are seen as mixtures of latent topics (modeled by the so-called *per-document topic distributions*), while topics are probability distributions over vocabulary words (modeled by *per-topic word distributions*).

These models have been originally designed to work with monolingual data, and they have been applied in monolingual contexts only. Following the ongoing growth of the World Wide Web and its omnipresence, users tend to abandon English as the universal language of the global network, since more and more content is available in their native languages. With the rapid development of

[1] This paper is not a novel work, but a short survey of our research in the field completed so far, along with concise definitions and modeling premises behind the concept of multilingual probabilistic topic modeling. Detailed descriptions of our work concerning both modeling and applications, along with extensive evaluations and comparisons with other methods (such as cross-lingual LSI [1, 2]) may be found in our published research papers, e.g. [3, 4, 5, 6, 7, 8].

Wikipedia and online social networks such as Facebook or Twitter, users have generated a huge volume of multilingual text resources. Multilingual probabilistic topic models (MuPTM-s) have recently emerged as a group of unsupervised, language-independent generative machine learning models that can be efficiently utilized on such large-volume non-parallel multilingual data. Due to its generic language-independent nature and the power of inference on unseen documents, these models have found many interesting applications. The knowledge from learned MuPTM-s has been used in many different cross-lingual tasks such as cross-lingual event clustering [3], cross-lingual document classification [4, 11], cross-lingual semantic similarity of words [12, 5, 8], cross-lingual information retrieval [6, 7] and others.

2 Multilingual Probabilistic Topic Modeling

2.1 Definitions and Assumptions

Assume that we are given a theme-aligned *multilingual corpus* $\mathcal{C}$ of $l = |\mathcal{L}|$ languages, where $\mathcal{L} = \{L_1, L_2, \dots, L_l\}$ is the set of languages. $\mathcal{C}$ is a set of text collections $\{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_l\}$ where each $\mathcal{C}_i = \{d^i_1, d^i_2, \dots, d^i_{dn_i}\}$ is a collection of documents in language L_i with vocabulary $V^i = \{w^i_1, w^i_2, \dots, w^i_{wn_i}\}$. Collections $\{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_l\}$ are *theme-aligned* if they discuss at least a portion of similar themes. Here, dn_i denotes the total number of documents in the corpus $\mathcal{C}_i$, while wn_i is the total number of words in V^i, and d^i_j denotes the j-th document in collection $\mathcal{C}_i$. We may now formally define *multilingual probabilistic topic modeling* and *cross-lingual latent topic extraction*.

Definition 1. Multilingual probabilistic topic model. A *multilingual probabilistic topic model* of a multilingual corpus $\mathcal{C}$ is a set of semantically coherent multinomial distributions of words with values $P_i(w^i|z_k)$, $i = 1, \dots, l$, for each vocabulary $V^1, \dots, V^i, \dots, V^l$ associated with text collections $\mathcal{C}_1, \dots, \mathcal{C}_i, \dots, \mathcal{C}_l \in \mathcal{C}$ given in languages $L_1, \dots, L_i, \dots, L_l$. w^i denotes a word from vocabulary V^i, and $P_i(w^i|z_k)$ is calculated for each $w^i \in V^i$. The probability scores $P_i(w^i|z_k)$ build *per-topic word distributions*, and they constitute a language-specific representation (e.g., a probability value is assigned only for words from V^i) of a language-independent latent cross-lingual concept, that is, latent cross-lingual topic $z_k \in \mathcal{Z}$. $\mathcal{Z} = \{z_1, \dots, z_K\}$ represents the set of all K latent cross-lingual topics present in the multilingual corpus. Each document in the multilingual corpus is thus considered a mixture of K latent cross-lingual topics from the set $\mathcal{Z}$. That mixture for some document $d^i \in \mathcal{C}_i$ is modeled by the probability scores $P_i(z_k|d^i)$ that build *per-document topic distributions*.

In summary, each document is represented as a mixture of latent cross-lingual topics, that is, each language-independent latent cross-lingual topic z_k has some probability to be found in a particular document (modeled by per-document topic distributions), and each cross-lingual topic has a language-specific representation in each language (modeled by language-specific per-topic word distributions). We can reinterpret the previous definition in the following way. Each cross-lingual topic from the set $\mathcal{Z}$ can be observed as a latent language-independent concept present in the multilingual corpus, but each language in the corpus uses only words from its own vocabulary to describe the content of that concept. For instance, having a multilingual collection in English, Italian and Dutch and discovering a topic on *Football*, that cross-lingual topic would be represented by words (actually probabilities over words) {*player, goal, coach,* …} in English, {*pallone (ball), calciatore (football player), squadra (team),* …} in Italian, and {*wedstrijd (match), elftal (football team), doelpunt (goal),* …} in Dutch. We have $\sum_{w^i \in V^i} P_i(w^i|z_k) = 1$, for each vocabulary V^i representing language L_i, and for each topic $z_k \in \mathcal{Z}$.

We say that a topic is *semantically coherent* if it assigns high probability scores to words that are semantically related. A desirable property of the cross-lingual topics learned from a theme-aligned multilingual corpus is to display both a strong *intra semantic coherence*, that is, words from the same vocabulary grouped together in the same topic are closely semantically related, as well as a strong *inter semantic coherence*, i.e., words across languages that represent the same cross-lingual topic are also closely semantically related.

Definition 2. Cross-lingual latent topic extraction. Given a theme-aligned multilingual corpus $\mathcal{C}$, the goal is to learn and extract a set $\mathcal{Z}$ of semantically coherent K latent cross-lingual topics $\{z_1, \dots, z_K\}$ that optimally describe the observed data, that is, the multilingual corpus $\mathcal{C}$. Extracting latent cross-lingual topics actually implies learning *per-document topic distributions* for each document in the corpus, and discovering language-specific representations of these latent topics given by *per-topic word distributions* in each language.

This shared and language-independent set of latent topics $\mathcal{Z}$ serves as the core of unsupervised cross-lingual knowledge transfer by means of MuPTM-s. It is the cross-lingual connection that bridges the gap across documents in different languages and transfers knowledge across languages in cases where translation resources and labeled instances are scarce or missing. The trained MuPTM-s may be inferred on unseen documents, where the *inference of the MuPTM* denotes learning per-document topic distributions for the new documents based on the training output.

2.2 A Very Short Overview of Current State-of-the-Art Models

We believe that bilingual LDA (see Appendix A), being the straightforward extension of the LDA model to the multilingual context, serves as a firm baseline for future advances in multilingual probabilistic topic modeling. However, although MuPTM is quite a novel concept, several other models have emerged over the last years. Current state-of-the-art MuPTM-s build upon the idea of standard pLSA and LDA, but they differ in the assumptions they make in their generative processes, and in knowledge that is presupposed before training (e.g., document alignment, prior word matchings or bilingual dictionaries). However, they all share the same concepts defined in Sect. 2.1, that is, the sets of output distributions, the set of latent cross-lingual topics that has to be discovered in a multilingual collection, etc.

Work from Zhao and Xing ([13, 14]) has focused on building topic models suitable for word alignment and statistical machine translation operations. Again inspired by monolingual LDA, they have designed several variants of topic models that operate on parallel corpora aligned at the sentence level. The topical structure at the level of aligned sentences or word pairs is used to re-estimate word translation probabilities and force alignments of words and phrases generated by the same topic. However, the growth of the global network and increasing amounts of comparable theme-aligned texts have formed a need for constructing more generic models that are applicable to such text collections. Standard probabilistic topic models coming from the families of pLSA and LDA cannot capture and accurately represent the structure of such theme-aligned multilingual text data in a form of joint cross-lingual topics. That inability comes from the fact that topic models rely on word co-occurrence information to group similar words into a single topic. In case of multilingual corpora (e.g., Wikipedia articles in English and Dutch) two related words in different languages will seldom co-occur in a monolingual text, and therefore these models are unable to group such pairs of words into a single coherent topic (see, e.g., [15, 16]). In order to anticipate that issue, there have been some efforts that trained monolingual probabilistic topic models on concatenated document pairs in two languages (e.g., [1, 17, 18, 19, 20, 21, 22]), but such approaches fail to build a shared latent cross-lingual topical space where the boundary between the topic representations with words in two languages is firmly established.

Recently, several novel models have been proposed that remove such deficiency. These models are trained on the individual documents in different languages and their output are joint cross-lingual topics in an aligned latent cross-lingual topical space. The BiLDA model [3] and its extensions to more than 2 languages ([12, 23]) constitute the current state-of-the-art in multilingual probabilistic topic modeling and have been validated in various cross-lingual tasks (e.g., [4, 11, 7]). These models require alignments at the document level *a priori* before training, which is easily obtained for Wikipedia or news articles. These document alignments provide hard links between topic-aligned semantically similar documents across languages.

Recently, there has been a growing interest in MuPTM from unaligned texts, again inspired by monolingual LDA. The MuTo model from Boyd-Graber and Blei [15] operates with *matchings* instead of words, where matchings consist of pairs of words that link words from the source vocabulary to words from the target vocabulary. These matchings are induced by the matching canonical correlation analysis (MCCA) [24] which ties together words with similar meanings across languages, where similarity is based on different features. Matchings are induced based on Pointwise Mutual Information (PMI) from parallel texts, machine-readable dictionaries and same orthographic features such as edit distance. A similar idea of using matchings has been investigated by Jagarlamudi and Daumé III [16]. In their JointLDA model, they also observe each topic as a mixture over these matchings (or *word concepts*, as they name them), where the matchings are acquired directly from a bilingual dictionary. Although these 2 models claim that they have removed the need for document alignment and are fit to mine topics from unaligned texts, they have introduced bilingual dictionaries as a new critical resource, These dictionaries have to be compiled from parallel data or hand-crafted, which is often more expensive and time-consuming than obtaining document alignments for Wikipedia or news data. Another work that aims to extract latent cross-lingual topics from unaligned datasets is presented by Zhang et al. [25]. Their Probabilistic Cross-lingual Latent Semantic Analysis (PCLSA) extends the common pLSA model [9] by regularizing its likelihood function with soft constraints defined by a bilingual dictionary. Similar to MuTo and JointLDA, a bilingual dictionary is a critical resource for PCLSA, since the dictionary-based constraints are the key to bridge the gap between languages by pushing related words in different vocabularies to occur in the same cross-lingual topics. The same relationship between pLSA and LDA [26] in the monolingual setting is also reflected between their multilingual extensions, PCLSA and BiLDA.

2.3 Applications of Multilingual Probabilistic Topic Models

As stressed before, all MuPTM-s revolve around two central sets of distributions: (1) per-document topic distributions and (2) per-topic word distributions. Discovered distributions can be directly employed to detect main themes discussed in texts, and to provide gists or summaries for large multilingual text collections. Per-document topic distributions for each document might be observed as a low-dimensional latent semantic representation of text in a new language-independent topic-document space, potentially better than the original word-based representation in some applications. In an analogous manner, since the number of topics is usually much lower than the number of documents in a collection, per-topic word distributions also model a sort of dimensionality reduction, as the original word-document space is transferred to a low-dimensional word-topic space. We can exploit these distributions and the power of inference of cross-lingual topic models on unseen documents in a number of cross-lingual tasks. Here, we list a few that we have been investigating in our research so far:

Cross-lingual event clustering. It refers to clustering of news stories written in different languages into groups of stories that describe the same event. An event can be observed as a mixture of different themes, where some themes are dominant while others are only marginally present. That phenomenon can be captured by cross-lingual topic models - per-document topic distributions will be higher for topics closely related to the themes prominent in a news story. Two news stories s_i and s_j are considered similar and are most likely discussing the same event if their per-document topic distributions are similar, that is, if the values $P(z_k|s_i)$ and $P(z_k|s_j)$ are similar for all $z_k \in \mathcal{Z}$, where $\mathcal{Z}$ is the set of K latent cross-lingual topics (see Sect. 2.1). Note that by utilizing the language-independent set $\mathcal{Z}$ and per-document topic distributions as the news story representation, we are able to perform the cross-lingual event clustering, i.e., the clustering of stories written in different languages. For more details, see the work of De Smet and Moens [3].

Cross-lingual document classification. The objective is to learn a classification model from the labeled documents in the source language and then apply it to the classification of documents in the target language. We can observe each document as a data instance and use the probabilities $P(z_k|d_j)$ from their per-document topic distributions as classification features. Again, by having the language independent set $\mathcal{Z}$ of K cross-lingual topics, we can operate in the same uniform feature space regardless of the specific languages in which documents were written. For more details, see the work of De Smet et al. [4].

Cross-lingual semantic word similarity. Since we have already detected that there exists a strong intra coherence and inter semantic coherence within cross-lingual topics, we could use per-topic word distributions for mining semantically similar words across languages. The similarity between two words can be computed based on the similarity between their conditional cross-lingual topic distributions, that is, two words w_1^S and w_2^T are semantically similar if the values $P(z_k|w_1^S)$ and $P(z_k|w_2^T)$ are similar for each $z_k \in \mathcal{Z}$. Semantically similar words may serve as a semantic cross-lingual lexicon which finds its application in tasks such as retrieval, text classification or machine translation. For more details, see the work of Vulić et al. [5, 8].

Cross-lingual information retrieval. It is easy to incorporate output from MuPTM-s into probabilistic language modeling framework for information retrieval. Each target document d_j^T can be presented as a mixture over cross-lingual topics from the set $\mathcal{Z}$ (see Sect. 2.1) as given by per-document topic distributions (with the values $P(z_k|d_j^T)$). Additionally, the values $P(w_i^S|z_k)$ from per-topic word distributions may be used to calculate the probability that a latent cross-lingual topic z_k will generate some source word w_i^S. If that word w_i^S is actually a word from the user's query written in the source language, the cross-lingual topics again serve as a bridge that links semantics of the query in the source language with semantics of the document written in the target language. MuPTM-s could be learned on one corpus and then inferred on another document collection that is used for retrieval. Also, semantically similar words across languages (see the previous application) may be integrated as useful additional evidences in cross-lingual retrieval models. For more details, see the work of Vulić et al. [6, 7].

References

[1] Susan T. Dumais, Thomas K. Landauer, and Michael Littman. Automatic cross-linguistic information retrieval using Latent Semantic Indexing. In *Proceedings of the SIGIR Workshop on Cross-Linguistic Information Retrieval*, pages 16–23, 1996.

[2] Thomas K. Landauer and Susan T. Dumais. Solutions to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[3] Wim De Smet and Marie-Francine Moens. Cross-language linking of news stories on the Web using interlingual topic modeling. In *Proceedings of the CIKM 2009 Workshop on Social Web Search and Mining (SWSM)*, pages 57–64, 2009.

[4] Wim De Smet, Jie Tang, and Marie-Francine Moens. Knowledge transfer across multilingual corpora via latent topics. In *Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 549–560, 2011.

[5] Ivan Vulić, Wim De Smet, and Marie-Francine Moens. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 479–484, 2011.

[6] Ivan Vulić, Wim De Smet, and Marie-Francine Moens. Cross-language information retrieval with latent topic models trained on a comparable corpus. In *Information Retrieval Technology - 7th Asia Information Retrieval Societies Conference (AIRS)*, pages 37–48, 2011.

[7] Ivan Vulić, Wim De Smet, and Marie-Francine Moens. Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval*, 2012.

[8] Ivan Vulić and Marie-Francine Moens. Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 449–459, 2012.

[9] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 50–57, 1999.

[10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, (3):993–1022, 2003.

[11] Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. Cross lingual text classification by mining multilingual topics from Wikipedia. In *Proceedings of the 4th International Conference on Web Search and Web Data Mining (WSDM)*, pages 375–384, 2011.

[12] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 880–889, 2009.

[13] Bing Zhao and Eric P. Xing. BiTAM: Bilingual topic admixture models for word alignment. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 969–976, 2006.

[14] Bing Zhao and Eric P. Xing. HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1689–1696, 2007.

[15] Jordan Boyd-Graber and David M. Blei. Multilingual topic models for unaligned text. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 75–82, 2009.

[16] Jagadeesh Jagarlamudi and Hal Daumé III. Extracting multilingual topics from unaligned comparable corpora. In *Proceedings of the 32th Annual European Conference on Advances in Information Retrieval (ECIR)*, pages 444–456, 2010.

[17] Michael Littman, Susan T. Dumais, and Thomas K. Landauer. In *Cross-Language Information Retrieval, chapter 5*, pages 51–62. Kluwer Academic Publishers, 1998.

[18] Jaime G. Carbonell, Jaime G. Yang, Robert E. Frederking, Ralf D. Brown, Yibing Geng, Danny Lee, Yiming Frederking, Robert E, Ralf D. Geng, and Yiming Yang. Translingual information retrieval: A comparative evaluation. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 708–714, 1997.

[19] Peter A. Chew, Brett W. Bader, Tamara G. Kolda, and Ahmed Abdelali. Cross-language information retrieval using PARAFAC2. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 143–152, 2007.

[20] Gui-Rong Xue, Wenyuan Dai, Qiang Yang, and Yong Yu. Topic-bridged pLSA for cross-domain text classification. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 627–634, 2008.

[21] Philipp Cimiano, Antje Schultz, Sergej Sizov, Philipp Sorg, and Steffen Staab. Explicit versus latent concept models for cross-language information retrieval. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence (IJCAI)*, pages 1513–1518, 2009.

[22] Benjamin Roth and Dietrich Klakow. Combining Wikipedia-based concept models for cross-language retrieval. In *Proceedings of the Information Retrieval Facility Conference (IRFC)*, pages 47–59, 2010.

[23] Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. Mining multilingual topics from Wikipedia. In *Proceedings of the 18th International World Wide Web Conference (WWW)*, pages 1155–1156, 2009.

[24] Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 771–779, 2008.

[25] Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. Cross-lingual latent topic extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1128–1137, 2010.

[26] Mark Girolami and Ata Kabán. On an equivalence between PLSI and LDA. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR)*, pages 433–434, 2003.

[27] Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. Topics in semantic representation. *Psychological Review*, 114(2):211–244, 2007.

[28] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[29] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7):424–440, 2007.

[30] Thomas P. Minka and John D. Lafferty. Expectation-propogation for the generative aspect model. In *Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence (UAI)*, pages 352–359, 2002.

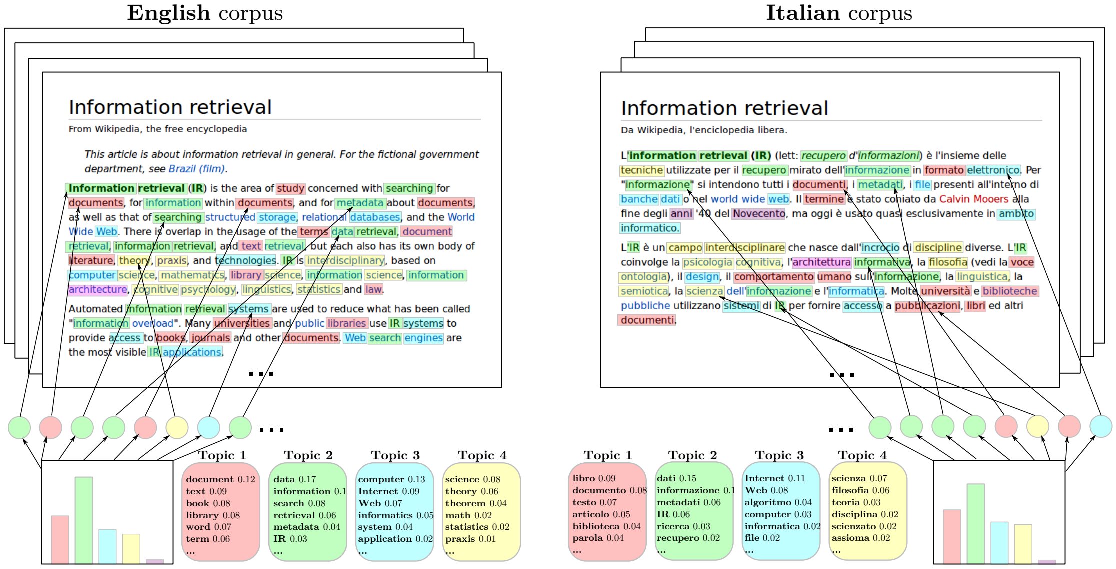
[31] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, pages 5228–5235, 2004.

[32] Yee Whye Teh and Dilan Görür. Indian buffet processes with power-law behavior. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1838–1846, 2009.

Appendix A: Bilingual LDA

The text in the appendix presents a comprehensive overview of the bilingual LDA model (BiLDA), which has been designed in our research group. In the paper we have defined the modeling assumptions and methodology of MuPTM from a high-level perspective, and here with bilingual LDA we present a case study of how the multilingual topic models work in practice, that is, how to train, infer and use these models.

Bilingual Latent Dirichlet Allocation (BiLDA) is a bilingual extension of the standard LDA model [10], tailored for modeling comparable bilingual document collections that are theme-aligned, but loosely equivalent to each other. An example of such a document collection is Wikipedia in 2 languages with paired articles.



Information retrieval

From Wikipedia, the free encyclopedia

This article is about information retrieval in general. For the fictional government department, see Brazil (film).

Information retrieval (**IR**) is the area of study concerned with searching for documents, for information within documents, and for metadata about documents, as well as that of searching structured storage, relational databases, and the World Wide Web. There is overlap in the usage of the terms data retrieval, document retrieval, information retrieval, and text retrieval, but each also has its own body of literature, theory, praxis, and technologies. IR is interdisciplinary, based on computer science, mathematics, library science, information science, information architecture, cognitive psychology, linguistics, statistics and law.

Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.

...

Information retrieval

Da Wikipedia, l'enciclopedia libera.

L'**information retrieval (IR)** (lett: *recupero d'informazioni*) è l'insieme delle tecniche utilizzate per il recupero mirato dell'informazione in formato elettronico. Per "informazione" si intendono tutti i documenti, i metadati, i file presenti all'interno di banche dati o nel world wide web. Il termine è stato coniato da Calvin Mooers alla fine degli anni '40 del Novecento, ma oggi è usato quasi esclusivamente in ambito informatico.

L'IR è un campo interdisciplinare che nasce dall'incrocio di discipline diverse. L'IR coinvolge la psicologia cognitiva, l'architettura informativa, la filosofia (vedi la voce ontologia), il design, il comportamento umano sull'informazione, la linguistica, la semiotica, la scienza dell'informazione e l'informatica. Molte università e biblioteche pubbliche utilizzano sistemi di IR per fornire accesso a pubblicazioni, libri ed altri documenti.

...

Figure 1: An illustrative overview of the intuitions behind MuPTM, and a high-level overview of the generative process for BiLDA. Histograms present the importance of each topic in each document, where some latent cross-lingual topics are more important for the particular document. Cross-lingual topics are latent language-independent concepts, but each language provides a language-specific interface to each cross-lingual topic (modeled by *per-topic word distributions*, presented by rounded rectangles). BiLDA assumes that each document pair is then generated as follows. First, choose the per-document topic distribution and, according to the distribution, for each word position choose a topic assignment (the colored circles). Following that, according to per-topic word distributions in that language, choose the specific word in the language from the corresponding cross-lingual topic that will occur at that word position. Documents that discuss similar themes tend to have similar distributions over cross-lingual topics, but when we operate in the multilingual context, different per topic-word distributions (the rounded rectangles) are used to generate the observed words in the documents. The generative process does not make any assumptions about word order as they appear in documents (the *bag-of-words assumption)*. The figure represents a toy example, and it is not based on real data.

BiLDA has been independently designed by several researchers ([23, 3, 12]). Unlike LDA, where each document is assumed to possess its own document-specific distribution over topics, the generative process for BiLDA assumes that each *document pair* shares the same distribution of topics. BiLDA can be observed as a three-level Bayesian network that models document pairs using a latent layer of shared topics. Fig. 1 visualizes the key intuitions behind the generative story of the BiLDA model, Fig. 2 shows its Bayesian structure in plate notation, while Alg. 2.1 presents its generative story. We assume that we are given a source language S and a target language T.

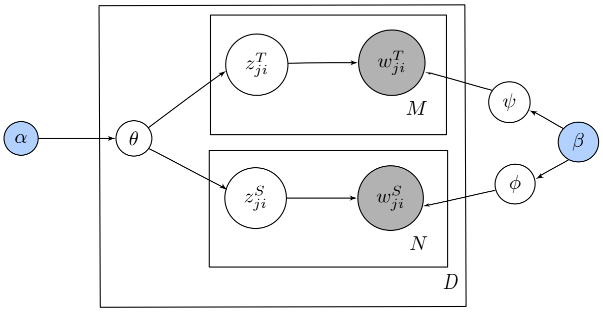


Figure 2: Graphical representation of the bilingual LDA (BiLDA) model. M and N denote lengths of the source document and the target document in terms of word tokens for each aligned document pair.

Algorithm 2.1: GENERATIVE STORY FOR BILDA()

```
initialize: (1) set the number of topics $K$;
(2) set values for Dirichlet priors $\alpha$ and $\beta$
sample $K$ times $\phi \sim Dirichlet(\beta)$
sample $K$ times $\psi \sim Dirichlet(\beta)$
for each document pair $d_j = \{d_j^S, d_j^T\}$
do
    sample $\theta_j \sim Dirichlet(\alpha)$
    for each word position $i \in d_j^S$
    do
        sample $z_{ji}^S \sim Multinomial(\theta)$
        sample $w_{ji}^S \sim Multinomial(\phi, z_{ji}^S)$
    for each word position $i \in d_j^T$
    do
        sample $z_{ji}^T \sim Multinomial(\theta)$
        sample $w_{ji}^T \sim Multinomial(\psi, z_{ji}^T)$
```

BiLDA takes advantage of the assumed thematic alignment at the level of linked documents by introducing a single variable θ shared by both documents. θ_j denotes the distribution of topics over each document pair d_j. For each document pair d_j, a per-document topic distribution[2] θ_j is sampled from a conjugate Dirichlet prior with K parameters $\alpha_1, \ldots, \alpha_K$. Then, with respect to θ_j, a topic z_{ji}^S is sampled. Each word w_{ji}^S at the position i in the source document of the current document pair d_j is then generated from a multinomial distribution $\phi_{z_{ji}^S}$. Similarly, each word w_{ji}^T of the target language is also sampled following the same procedure. Note that words at the same positions for source and target documents in a document pair need not be sampled from the same cross-lingual topic. The only constraint imposed by the model is that the overall distributions of topics over documents in a document pair modeled by θ_j have to be the same. In practice, it does not pose a problem when dealing with comparable data such as Wikipedia articles.

According to Griffiths et al. [27], each hyper-parameter α_j could be interpreted as a prior observation count for the number of times topic j is sampled in a document before having observed any actual words. If we do not possess any prior knowledge about themes in a text collection, it is reasonable to assume that all topics are a priori equally likely. Therefore, it is convenient to use a symmetric Dirichlet distribution with a single hyper-parameter α such that $\alpha_1 = \ldots = \alpha_K = \alpha$. Similarly, a symmetric Dirichlet prior is placed on ϕ and ψ with a single hyper-parameter β. β can be interpreted as a prior observation count on the number of times words in each language are sampled from a topic before any observations of actual words. Placing these Dirichlet prior distributions on multinomial distributions θ, ϕ and ψ results in smoothed per-topic word and per-document topic distributions, where the values for α and β determine the degree of smoothing.

[2] The correct term here should be per-pair topic distribution for BiLDA and per-tuple topic distribution in case when more than 2 languages are involved, but we have decided to keep the original name of the distribution in order to draw a direct comparison with standard monolingual LDA.

A natural extension of BiLDA that operates with more than 2 languages, called *polylingual topic model* is presented by Mimno et al. [12]. A similar model is proposed by Ni et al. [23]. Instead of document pairs, they deal with *document tuples* (where links between documents in a tuple are given), but the assumptions made by their model remain the same.

Estimating the Bilingual LDA Model

The goal of training the BiLDA model is to discover the layer of latent cross-lingual topics that describe observed data, i.e., a given bilingual document collection in an optimal way. It means that the most likely values for θ, ϕ and ψ have to be found by the training procedure. In simple words, we need to detect and learn which words are important for a particular topic in each language (that is reflected in per-topic word distributions ϕ and ψ), and which topics are important for a particular document pair (as reflected in per-document topic distribution θ). Similarly to the LDA model, topic discovery for BiLDA is complex and cannot be solved by an exact learning procedure.

There exist a few approximate learning techniques. Variational estimation for the monolingual LDA was used as the estimation technique in the seminal paper by Blei et al. [10]. Other estimation techniques for the monolingual case include Gibbs sampling [28, 29], and expectation propagation [30, 31]. Teh and Görür [32] study the influence of different estimation techniques on the quality of monolingual LDA (evaluated by perplexity) and show that optimization of the hyper-parameters leads to the stable performance of LDA, no matter what estimation technique is used for training. An extension of the variational method to the multilingual context and its complete formulation for BiLDA was also proposed and described by the authors [3]. Due to its prevalent use in probabilistic topic modeling literature in both monolingual and multilingual contexts [15, 12, 16, 5], we here provide a short overview of Gibbs sampling as the estimation technique. A more detailed overview of Gibbs sampling for BiLDA is also provided by the authors [7].

The Gibbs sampling procedure for BiLDA requires two sets of formulas to converge to correct distributions: one for each topic z^S_{ji} (a topic assigned to a word position i that generated word w^S_{ji} in a document pair d_j), and one for each topic z^T_{ji}. θ, ψ and ϕ are not calculated directly, but estimated afterwards. Therefore, they are integrated out of all the calculations, which actually leaves z^S_{ji}-s and z^T_{ji}-s as the only hidden variables. For the source part S of each document pair d_j and each word position i, the probability is calculated that z^S_{ji} assumes, as its new values, one of the K possible topic indices (from a set of K topics), as indicated by variable k:

$$\begin{aligned}\text{sample } z^S_{ji} &\sim P(z^S_{ji} = k|\mathbf{z}^S_{\neg ji}, \mathbf{z}^T_j, \mathbf{w}^S_j, \mathbf{w}^T_j, \alpha, \beta)\\ &\sim \int_{\theta_j}\int_{\phi} P(z^S_{ji} = k|, \mathbf{z}^S_{\neg ji}, \mathbf{z}^T_j, \mathbf{w}^S_j, \mathbf{w}^T_j, \alpha, \beta, \theta_j, \phi) d\phi d\theta_j\end{aligned}$$

In this formula, $\mathbf{z}^T_j$ refers to all target topic indices for document pair d_j, and $\mathbf{z}^S_{\neg ji}$ denotes all source topic indices in d_j excluding z^S_{ji}. $\mathbf{w}^S_j$ denotes all source words, $\mathbf{w}^T_j$ all target words in the current document pair d_j. Sampling for the target side T is done in an analogical manner. Due to its close resemblance to the Gibbs sampling procedure for monolingual LDA, we omit all the exact derivation steps (see, e.g., [7]). The final updating formulas for Gibbs sampling for BiLDA are:

$$P(z^S_{ji} = k) \propto \frac{n^S_{j,k,\neg i} + n^T_{j,k} + \alpha}{n^S_{j,\cdot,\neg i} + n^T_{j,\cdot} + K \cdot \alpha} \cdot \frac{v^S_{k,w^S_{ji},\neg} + \beta}{v^S_{k,\cdot,\neg} + |V^S| \cdot \beta} \tag{1}$$

$$P(z^T_{ji} = k) \propto \frac{n^T_{j,k,\neg i} + n^S_{j,k} + \alpha}{n^T_{j,\cdot,\neg i} + n^S_{j,\cdot} + K \cdot \alpha} \cdot \frac{v^T_{k,w^T_{ji},\neg} + \beta}{v^T_{k,\cdot,\neg} + |V^T| \cdot \beta}. \tag{2}$$

The counter variable $n^S_{j,k}$ counts the number of times that topic with index k has been sampled from the multinomial distribution specific to document d^S_j from the document pair d_j. $n^S_{j,k,\neg i}$ has the same meaning, except that the current z^S_{ji} is not counted. Another counter variable, $v^S_{k,w^S_{ji},\neg}$ counts the number of times w^S_{ji} has been generated by topic k, but not counting the current w^S_{ji}, i.e., it is $v^S_{k,w_{ji}} - 1$. In these counters a dot $(\cdot)$ denotes summation over all values of the variable whose index the dot takes, that is, all topics in case of $n^S_{j,\cdot}$ and all words in $v^S_{k,\cdot,\neg}$. The meaning of the counter variables for the target side T is the same. $|V^S|$ and $|V^T|$ are vocabulary sizes for the source and the target language, respectively.

With formulas (1) and (2). each z^S_{ji} and z^T_{ji} of each document pair is sampled and cyclically updated. After a random initialization, usually using a uniform distribution, the sampled values will converge to samples taken from the real joint distribution of θ, ϕ and ψ, after a time called the *burn-in* period. From one complete burned-in Gibbs sample of the whole document collection, the per-topic word distributions and per-document topic distributions are estimated.

Per-document topic distributions provide distributions of latent cross-lingual topics for each document in a collection. They reveal how important each topic is for a particular document. First, we need to establish the exact formula for per-document topic distributions for documents in an aligned document pair using Eq. (1) and Eq. (2):

$$P(z_k|d_j) = \theta_{j,k} = \frac{n^S_{j,k} + n^T_{j,k} + \alpha}{\sum_{k^*=1}^{K} n^S_{j,k^*} + \sum_{k^*=1}^{K} n^T_{j,k^*} + K\alpha} \tag{3}$$

Per-topic word distributions measure the importance of each word in each language for a particular cross-lingual topic z_k. Given a source language with vocabulary V^S, and a target language with vocabulary V^T, and following Eq. (1), a probability that some word $w^S_i \in V^S$ will be generated by the cross-lingual topic z_k is given by:

$$P(w^S_i|z_k) = \phi^S_{k,i} = \frac{v^S_{k,w^S_i} + \beta}{\sum_{i^*=1}^{|V^S|} v^S_{k,w^S_{i^*}} + |V^S|\beta} \tag{4}$$

The same formula, but now derived from Eq. (2) is used for the per-topic word distributions (ψ) for the target language.

2.3.1 Output of the Model

Table 1: Randomly selected examples of latent cross-lingual topics represented by top 10 words based on their counts after Gibbs sampling. Topics are discovered by BiLDA trained on Wikipedia for various language pairs: French-English (FR-EN), Dutch-English (NL-EN), Italian-English (IT-EN), and Spanish-English (ES-EN). For non-English words we have provided corresponding English translations. K=100 for all models.

FR-EN Topic 17	NL-EN Topic 55	IT-EN Topic 73	ES-EN Topic 52
moteur (engine)	gebouw (building)	rete (network)	dinero (money)
voiture (vehicle)	eeuw (century)	chiave (key)	mercado (market)
automobile (car)	meter (meter)	protocollo (protocol)	precio (price)
vitesse (speed)	kasteel (castle)	server (server)	bienes (goods)
constructeur (constructor)	bisschop (bishop)	messaggio (message)	valor (value)
roue (wheel)	stad (city)	connessione (connection)	cantidad (amount)
vapeur (steam)	gebouwd (built)	client (client)	oferta (offer)
puissance (power)	theater (theater)	servizion (service)	pago (payment)
diesel (diesel)	museum (museum)	indirizzo (address)	impuesto (tax)
cylindre (cylinder)	tuin (garden)	sicurezza (security)	empresa (company)
engine	building	link	economic
car	court	network	price
vehicle	built	display	money
fuel	garden	calendar	market
speed	museum	client	capital
power	palace	key	tax
production	construction	server	goods
design	theater	protocol	interest
diesel	tower	address	demand
drive	castle	packet	inflation

Since the model possesses a fully generative semantics, it is possible to train the model on one multilingual corpus (e.g., Wikipedia) and then infer it on some other, previously unseen corpus. Inferring a model on a new corpus means calculating per-document topic distributions for the unseen documents based on the output of the trained model. Inference on the unseen documents is performed on only one language at a time, e.g., if we train on English-Dutch Wikipedia, we can use the trained BiLDA model to learn per-document topic distributions for Dutch news stories, and then separately for English news stories. In short, we again randomly sample and then iteratively update topic assignments for each word position in an unseen document, but then use the fixed v counters learned in training to update the topic assignments (the n counters). Since the inference is performed monolingually, dependencies on the topic assignments from another language are removed from the updating formulas. Hence, similar to Eq. (1), the updating formula for the source language is:

$$P(z^S_{ji} = k) \propto \frac{n^S_{j,k,\neg i} + \alpha}{n^S_{j,\cdot,\neg i} + K \cdot \alpha} \cdot \frac{v^S_{k,w^S_{ji}} + \beta}{v^S_{k,\cdot} + |V^S| \cdot \beta} \qquad (5)$$

Learning a MuPTM on one multilingual corpus and then inferring that model on previously unseen data constitutes the key concept of cross-lingual *knowledge transfer* by means of MuPTM-s.

Another way of looking at output of a probabilistic topic model is by simply inspecting top words associated with a particular topic learned during training. It is much easier for humans to judge semantic coherence of latent cross-lingual topics and their alignment across languages when observing the actual words constituting a topic. These words provide a shallow qualitative representation of the latent topic space, and could be seen as direct and comprehensive word-based summaries of a large document collection. In other words, humans can get the first clue “what all this text is about in the first place”. Some latent cross-lingual topics extracted by BiLDA trained on aligned Wikipedia articles are provided in Table 1. We observe strong intra semantic coherence as well as strong inter semantic coherence.