Information Processing and Management 51 (2015) 111–147

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications

Ivan Vulić [a,*], Wim De Smet [a], Jie Tang [b], Marie-Francine Moens [a]

[a] Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium
[b] Department of Computer Science and Technology, Tsinghua University, 100084 Beijing, China

ARTICLE INFO

Article history:
Received 4 February 2013
Received in revised form 11 August 2014
Accepted 18 August 2014
Available online 7 October 2014

Keywords:
Multilingual probabilistic topic models
Cross-lingual text mining
Cross-lingual knowledge transfer
Cross-lingual information retrieval
Language-independent data representation
Non-parallel data

ABSTRACT

Probabilistic topic models are unsupervised generative models which model document content as a two-step generation process, that is, documents are observed as mixtures of latent concepts or 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 multilingual probabilistic topic modeling (MuPTM) and present the first full overview of the current research, methodology, advantages and limitations in MuPTM. As a representative example, we choose a natural extension of the omnipresent LDA model to multilingual settings called bilingual LDA (BiLDA). We provide a thorough overview of this representative multilingual model from its high-level modeling assumptions down to its mathematical foundations. We demonstrate how to use the data representation by means of output sets of (i) per-topic word distributions and (ii) per-document topic distributions coming from a multilingual probabilistic topic model in various real-life cross-lingual tasks involving different languages, without any external language pair dependent translation resource: (1) cross-lingual event-centered news clustering, (2) cross-lingual document classification, (3) cross-lingual semantic similarity, and (4) cross-lingual information retrieval. We also briefly review several other applications present in the relevant literature, and introduce and illustrate two related modeling concepts: topic smoothing and topic pruning. In summary, this article encompasses the current research in multilingual probabilistic topic modeling. By presenting a series of potential applications, we reveal the importance of the language-independent and language pair independent data representations by means of MuPTM. We provide clear directions for future research in the field by providing a systematic overview of how to link and transfer aspect knowledge across corpora written in different languages via the shared space of latent cross-lingual topics, that is, how to effectively employ learned per-topic word distributions and per-document topic distributions of any multilingual probabilistic topic model in various cross-lingual applications.

* Corresponding author. Tel.: +32 16 32 87 14.
E-mail addresses: ivan.vulic@cs.kuleuven.be (I. Vulić), wdesmet@gmail.com (W. De Smet), jie.tang@tsinghua.cn.edu (J. Tang), marie-francine.moens@cs.kuleuven.be (M.-F. Moens).

http://dx.doi.org/10.1016/j.ipm.2014.08.003
0306-4573/

1. Introduction

Probabilistic latent topic models such as probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999b, 1999a) and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003b) 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 probabilistic latent topic models are all based upon the idea that there exist latent variables, that is, *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, while topics are simply probability distributions over vocabulary words. A topic representation of a document constitutes a high-level language-independent view of its content, unhindered by a specific word choice, and it improves on text representations that contain synonymous or polysemous words (Griffiths, Steyvers, & Tenenbaum, 2007).

Probabilistic topic modeling constitutes a very general framework for unsupervised topic mining, and over the years it has been employed in miscellaneous tasks in a wide variety of research domains, e.g., for object recognition in computer vision (e.g., Li & Perona, 2005; Russell, Freeman, Efros, Sivic, & Zisserman, 2006; Wang & Grimson, 2007), dialogue segmentation (e.g., Purver, Körding, Griffiths, & Tenenbaum, 2006), video analysis (e.g., Wang, Ma, & Grimson, 2009), automatic harmonic analysis in music (e.g., Arenas-Garca et al., 2007; Hu & Saul, 2009), genetics (e.g., Blei, Franks, Jordan, & Mian, 2006), and others.

Being originally proposed for textual data, probabilistic topic models have also organically found many applications in natural language processing (NLP). Discovered distributions of words over topics (further *per-topic word distributions*) and distributions of topics over documents (further *per-document topic distributions*) can be directly employed to detect main themes[1] discussed in texts, and to provide gists or summaries for large text collections (see, e.g., Hofmann, 1999b; Blei et al., 2003b; Griffiths & Steyvers, 2004; Griffiths et al., 2007). Per-document topic distributions for each document might be observed as a low-dimensional latent semantic representation of text in a new 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 lower-dimensional word-topic space. Apart from the straightforward utilization of probabilistic topic models as direct summaries of large document collections, these two sets of probability distributions have been utilized in a myriad of NLP tasks, e.g., for inferring captions for images (Blei & Jordan, 2003), sentiment analysis (e.g., Mei, Ling, Wondra, Su, & Zhai, 2007; Titov & McDonald, 2008), analyzing topic trends for different time intervals in scientific literature, social networks and e-mails (e.g., Wang & McCallum, 2006; McCallum, Wang, & Corrada-Emmanuel, 2007; Hall, Jurafsky, & Manning, 2008), language modeling in information retrieval (e.g., Wei & Croft, 2006; Yi & Allan, 2009), document classification (e.g., Blei et al., 2003b; Lacoste-Julien, Sha, & Jordan, 2008), word sense disambiguation (e.g., Boyd-Graber, Blei, & Zhu, 2007), modeling distributional similarity of terms (e.g., Ritter, Mausam, & Etzioni, 2010; Dinu & Lapata, 2010), etc. Lu, Mei, and Zhai, 2011 examine task performance of pLSA and LDA as representative monolingual topic models in typical tasks of document clustering, text categorization and ad hoc information retrieval. Data representation, i.e., representations of words and documents in all applications presented in this article will be based on those per-topic word distributions and per-document topic distributions.

However, all these models have been 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 in today's increasingly connected world, users tend to abandon English as the *lingua franca* of the global network, since more and more content becomes available in their native languages or even dialects and different community languages (e.g., the idiomatic usage of the same language typically differs between scientists, social media consumers or the legislative domain). It is difficult to determine the exact number of languages in the world, but the estimations vary between 6000 and 7000 languages and almost 40,000 unofficial languages and dialects.[2] It is extremely time-consuming and labor-intensive to build quality *translation resources* and *parallel corpora* for each single language/dialect pair. Therefore, we observe an increasing interest in language-independent unsupervised corpus-based cross-lingual text mining from non-parallel corpora without any additional translation resources. High-quality parallel corpora where documents are sentence-aligned exact translations of each other (such as Europarl (Koehn, 2005)) are available only for a restricted number of languages and domains. There has been a recent interest to build parallel corpora from the Web (e.g., Resnik & Smith, 2003; Munteanu & Marcu, 2005, 2006), but the obtained parallel data still typically remain of limited size and scope as well as domain-restricted (e.g., parliamentary proceedings).

With the rapid development of Wikipedia and online social networks such as Facebook or Twitter, users have generated a huge volume of multilingual text resources. The user-generated data are often noisy and unstructured, and seldom well-paired across languages. However, unlike parallel corpora, such *comparable corpora*, where texts in one language are paired with texts in another language discussing the same themes or subjects, are abundant in various online sources (e.g., Wikipedia or news sites). Documents from comparable corpora do not necessarily share all their themes with their counterparts in the other language, but, for instance, Wikipedia articles discussing the same subject, or news stories discussing the

[1] To avoid confusion, we talk about *themes* when we address the true content of a document, while we talk about topics when we address the probability distributions constituting a topic model.

[2] Source: http://www.ethnologue.com.