Monolingual and Cross-Lingual Probabilistic Topic Models and Their Applications in Information Retrieval

Marie-Francine Moens and Ivan Vulić

KU Leuven,
Department of Computer Science,
Celestijnenlaan 200A,
B-3001 Heverlee, Belgium
{marie-francine.moens,ivan.vulic}@cs.kuleuven.be

Abstract. Probabilistic topic models are a group of unsupervised generative machine learning models that can be effectively trained on large text collections. They 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 multilingual data (e.g., Wikipedia or news data discussing the same events). Probabilistic topics models offer an elegant way to represent content across different languages. Their probabilistic framework allows for their easy integration into a language modeling framework for monolingual and cross-lingual information retrieval. Moreover, we present how to use the knowledge from the topic models in the tasks of cross-lingual event clustering, cross-lingual document classification and the detection of cross-lingual semantic similarity of words. The tutorial also demonstrates how semantically similar words across languages are integrated as useful additional evidences in cross-lingual information retrieval models.

Keywords: Probabilistic topic models, Cross-lingual retrieval, Ranking models, Cross-lingual text mining.

1 Goals and Scope

This *half day tutorial* presents an overview of recent advances in probabilistic topic modeling, which includes: (1) A high-level overview of the key intuitions and assumptions behind probabilistic topic modeling in general and cross-lingual probabilistic topic modeling in specific; (2) The methodology and mathematical foundations; and (3) The application of these probabilistic topic models in various monolingual and cross-lingual tasks, with a special focus on monolingual and cross-lingual retrieval models.

The tutorial first introduces the concept of probabilistic topic modeling, starting from monolingual contexts, where we introduce the key intuitions and describe the most prominent monolingual models such as pLSA and LDA. We provide a short overview of their modeling premises, training and inference procedures, and potential applications (e.g., language modeling for ad-hoc retrieval).

Following that, we gradually shift to the cross-lingual domain. We present a representative cross-lingual topic model called bilingual LDA (BiLDA) developed in our research group, which is designed as a natural extension of the monolingual LDA model, but is able to operate with multilingual comparable data (e.g., Wikipedia articles and news stories that discuss similar events). We again go deeper into the modeling assumptions of BiLDA. We explain its generative story, its training techniques (variational inference and Gibbs sampling) and its inference procedure on unseen text documents. We provide insight into other cross-lingual probabilistic topic models developed in recent years. We also present how to model more complex cross-lingual topic models that aim to catch more fine-grained structures and redundancies in training data.

Finally, an important part of the tutorial focuses on the applications of the monolingual and cross-lingual topic models (we use BiLDA in all cross-lingual applications). The emphasis is on retrieval models. We explain in detail how to integrate the knowledge from probabilistic topic models into probabilistic retrieval models in both monolingual and cross-lingual settings. Besides that application, we also present how to use the knowledge from the models for the tasks of cross-lingual event clustering, cross-lingual document classification and the detection of cross-lingual semantic similarity of words. The tutorial presents how such semantically similar words across languages are integrated as useful additional evidences in cross-lingual retrieval models.

The tutorial's main goal is to provide the participants a clear, comprehensible and systematic overview of probabilistic topic modeling with a special focus on its cross-lingual version, and the integration of the functionality of such models into probabilistic models for various monolingual and cross-lingual tasks, with a special focus on retrieval models. A small set of interactive demos and short exercises with examples will help the audience grasp the key concepts. By attending the tutorial, attendants will:

- Acquire an understanding of basic concepts of probabilistic topic modeling in general, as well as cross-lingual topic modeling in specific.
- Get acquainted with the current state-of-the-art research in the field of monolingual and cross-lingual probabilistic topic modeling that could serve as a valuable starting point for further research.
- Understand how to train and infer probabilistic topic models, with a special focus on multilingual comparable training data.
- Understand how to represent documents in a uniform way regardless of their actual language by means of probabilistic topic models.
- Come to know how to integrate the knowledge from probabilistic topic models into probabilistic models for information retrieval in both monolingual and cross-lingual settings.
- Understand how to effectively exploit the functionality of cross-lingual topic models in a variety of cross-lingual tasks.
- Be able to use the knowledge from probabilistic topic models for building novel retrieval models, and be able to effectively utilize the knowledge from topic models in other applications which are not mentioned in the tutorial.

The tutorial covers the field of probabilistic topic modeling in monolingual and multilingual settings from a broad perspective, so researchers with different levels of prior knowledge are welcome. The audience which is not entirely familiar with the concept of

probabilistic topic modeling will gain an invaluable didactic insight, as the tutorial systematically covers the topic from the very foundations and mathematical background to the actual practical use. The audience more interested in practical IR models will learn how the current state-of-the-art topic-based retrieval models work, and how to exploit the power of probabilistic topic modeling in information retrieval.

2 Tutorial Overview

The tutorial comprises the following parts:

1. *Key Intuitions and Methodology of Probabilistic Topic Modeling*: Situating the work on probabilistic topic modeling. Introduction to unsupervised generative models, basic probabilistic concepts and a high-level overview of probabilistic topic modeling. (Moens)
2. *Mathematics of Probabilistic Topic Modeling*: pLSA and LDA models and their generative stories. Dirichlet priors. Training and inference (variational Dirichlet, Gibbs sampling). Output per-document topic distributions and per-topic word distributions. Determining the number of topics. Evaluating the models. Applications of monolingual topic models - clustering, text classification and ad-hoc information retrieval. (Moens)
3. *Probabilistic Topic Modeling and Multilinguality*: Moving to the multilingual context. An overview of data resources. Web data and multilinguality. Different corpora types. Problems with translation resources for cross-lingual tasks. (Vulić)
4. *Probabilistic Topic Modeling in Multilingual Settings*: Defining the space of cross-lingual topics and cross-lingual topic extraction. Bilingual LDA. Training and inference (variational Dirichlet, Gibbs sampling). Language-independent uniform document representations. Language-independent per-document topic distributions. Language-specific per-topic word distributions. Other probabilistic topic models tailored for multilingual settings. (Moens)
5. *Cross-Lingual Probabilistic Topic Models in Practice*: Exploiting the knowledge from the output per-topic word and per-document topic distributions in various tasks - cross-lingual event clustering, cross-lingual document classification, cross-lingual semantic word similarity and translation mining. Cross-lingual topic models in cross-lingual information retrieval. Using the knowledge from topic models for query expansion. Combining evidences from topic models in retrieval models. Evaluation. (Vulić)

Each of the five parts should take about 30-40 minutes, and breaks will be put in between. In the final part, a few interactive exercises and a discussion is foreseen where the participants will be able to apply the knowledge gained.

The tutorial is given in the framework of the EU ICT COST Action MUMIA (Multilingual and Multifaceted Interactive information Access - IC1002).

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022 (2003)
2. Boyd-Graber, J.L., Blei, D.M.: Multilingual topic models for unaligned text. In: *Proceedings of UAI*, pp. 75–82 (2009)

3. Cimiano, P., Schultz, A., Sizov, S., Sorg, P., Staab, S.: Explicit versus latent concept models for cross-language information retrieval. In: *Proceedings of IJCAI*, pp. 1513–1518 (2009)
4. De Smet, W., Moens, M.-F.: Cross-language linking of news stories on the Web using interlingual topic models. In: *Proceedings of SWSM 2009*, pp. 57–64 (2009)
5. De Smet, W., Tang, J., Moens, M.-F.: Knowledge Transfer across Multilingual Corpora via Latent Topics. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) *PAKDD 2011, Part I. LNCS*, vol. 6634, pp. 549–560. Springer, Heidelberg (2011)
6. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6(6), 721–741 (1984)
7. Griffiths, T.L., Steyvers, M., Tenenbaum, J.B.: Topics in semantic representation. *Psychological Review* 114(2), 211–244 (2007)
8. Hofmann, T.: Probabilistic Latent Semantic Indexing. In: *Proceedings of SIGIR*, pp. 50–57 (1999)
9. Jagarlamudi, J., Daumé III, H.: Extracting Multilingual Topics from Unaligned Comparable Corpora. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) *ECIR 2010. LNCS*, vol. 5993, pp. 444–456. Springer, Heidelberg (2010)
10. Lavrenko, V., Choquette, M., Croft, W.B.: Cross-lingual relevance models. In: *Proceedings of SIGIR*, pp. 175–182 (2002)
11. Lu, Y., Mei, Q., Zhai, C.: Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval* 14(2), 178–203 (2011)
12. Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A.: Polylingual topic models. In: *Proceedings of EMNLP*, pp. 880–889 (2009)
13. Ni, X., Sun, J.-T., Hu, J., Chen, Z.: Cross-lingual text classification by mining multilingual topics from Wikipedia. In: *Proceedings of WSDM*, pp. 375–384 (2011)
14. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: *Proceedings of SIGIR*, pp. 275–281 (2001)
15. Roth, B., Klakow, D.: Combining Wikipedia-Based Concept Models for Cross-Language Retrieval. In: Cunningham, H., Hanbury, A., Rüger, S. (eds.) *IRFC 2010. LNCS*, vol. 6107, pp. 47–59. Springer, Heidelberg (2010)
16. Vulić, I., De Smet, W., Moens, M.-F.: Identifying word translations from comparable corpora using latent topic models. In: *Proceedings of ACL*, pp. 479–484 (2011)
17. Vulić, I., De Smet, W., Moens, M.-F.: Cross-Language Information Retrieval with Latent Topic Models Trained on a Comparable Corpus. In: Salem, M.V.M., Shaalan, K., Oroumchian, F., Shakeri, A., Khelalfa, H. (eds.) *AIRS 2011. LNCS*, vol. 7097, pp. 37–48. Springer, Heidelberg (2011)
18. Vulić, I., De Smet, W., Moens, M.-F.: Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. In: *Information Retrieval* (2012)
19. Vulić, I., Moens, M.-F.: Detecting highly confident word translations from comparable corpora without any prior knowledge. In: *Proceedings of EACL*, pp. 449–459 (2012)
20. Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D.: Evaluation methods for topic models. In: *Proceedings of ICML*, pp. 1105–1112 (2009)
21. Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: *Proceedings of SIGIR*, pp. 178–185 (2006)
22. Yi, X., Allan, J.: A Comparative Study of Utilizing Topic Models for Information Retrieval. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) *ECIR 2009. LNCS*, vol. 5478, pp. 29–41. Springer, Heidelberg (2009)
23. Zhang, D., Mei, Q., Zhai, C.: Cross-lingual latent topic extraction. In: *Proceedings of ACL*, pp. 1128–1137 (2010)