Extrinsic Evaluation of Topic Models on Unknown Corpora

Carsten Schnober, Iryna Gurevych

Ubiquitous Knowledge Processing Lab (UKP)
Department of Computer Science, Technische Universität Darmstadt (TUDA) /
German Institute for International Educational Research (DIPF)

LDA topic modelling (Blei et al., 2003) has become a popular tool for corpus analysis (Griffiths & Steyvers, 2004; Hall et al., 2008; Templeton et al., 2011). Its unsupervised nature makes LDA appealing for the analysis of newly digitized corpora: human-interpretable topics and their distributions can be inferred without requiring previous knowledge. In practice, however, this also makes it difficult for users to build trust in the results, and to evaluate different parameterizations.

We will present our practical experiences from using information retrieval (IR) as an extrinsic task for which, amongst many others, LDA can be applied (Wei et al., 2006). This allows us to use established evaluation metrics, precision and recall, in a realistic use case. We assume that a model that performs well on IR fits related tasks as well.

We tackle the following challenge: how can we measure precision and recall on a large corpus we know nothing about? Precision is the easier of the two parts: after submitting a query human experts manually count the portion of relevant documents among the first n hits (Prec@n).

It is unfeasible, however, to measure the portion of all relevant documents in the result set – the recall – because their total number is unknown. Therefore, we present a method to approximate the recall by extrapolating from smaller sub-corpora. Traditionally, this is solved by random pooling; exploiting available expert knowledge yields more targeted extrapolations.

We semi-automatically compile sub-corpora that represent a sub-topic of a more general super-topic. The following example sub-topics of 'modern infrastructure' stem from the digital humanities project "Children and their World"[1]:

1. Suez Canal
2. Kiel Canal

We now look manually for one or more sub-topics that can be described by terms that

a) occur in every relevant document
b) do not occur in any irrelevant document

For sub-topic 1, our experts find that 'Suez' complies with both conditions:

[1] "Children and their World": http://welt-der-kinder.gei.de/

a) In our data, every document that discusses the Suez Canal should mention 'Suez'.
b) In our data, no document that is irrelevant for the Suez Canal should mention 'Suez'.

With a term search for 'Suez', we can thus automatically create a Suez Canal sub-corpus that contains all relevant documents and no irrelevant ones. Random checks confirm the qualitative analysis by the domain experts.

Sub-topic 2 serves as an example for an unexploitable sub-topic: we expect the German city 'Kiel' to also occur in documents that do not discuss the Kiel Canal; hence, condition b) is violated. Note that if we are unable to find a compliant sub-topic, the presented method cannot be applied; established pooling methods may serve as a fall-back solution.

To evaluate a model's performance on a query, we first count the Prec@n score in the result set. For the recall, we measure the portion of documents from the Suez Canal sub-corpus in the same result set and repeat the procedure for other sub-corpora. Eventually, we aggregate these sub-recalls to extrapolate the score for the full query.

References

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

Griffiths, Thomas L., and Mark Steyvers. "Finding Scientific Topics." Proceedings of the National Academy of Sciences of the United States of America 101, no. Suppl 1 (2004): 5228–35. doi:10.1073/pnas.0307752101.

Hall, David, Daniel Jurafsky, and Christopher D. Manning. "Studying the History of Ideas Using Topic Models." In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 363–71. Association for Computational Linguistics, 2008. http://dl.acm.org/citation.cfm?id=1613763.

Templeton, Clay, Travis Brown, Sayan Battacharyya, and Jordan Boyd-Graber. "Mining the Dispatch under Supervision: Using Casualty Counts to Guide Topics from the Richmond Daily Dispatch Corpus." In Chicago Colloquium on Digital Humanities and Computer Science. Chicago, Illinois, USA, 2011. http://www.umiacs.umd.edu/~jbg/docs/slda_civil_war.pdf.

Wei, Xing, and W. Bruce Croft. "LDA-Based Document Models for Ad-Hoc Retrieval." In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 178–85. SIGIR '06. New York, NY, USA: ACM, 2006. doi:10.1145/1148170.1148204.