Hurvitz, A. (2013). *Late Biblical Hebrew*, Khan.
Khan, G. (ed.) (2013). *Encyclopedia of Hebrew Language and Linguistics*, Vol. 4, Leiden, Brill, 2013.
Kutscher, E. Y. (1974). *The Language and Linguistic Background of the Isaiah Scroll (1QIsaa),* STDJ 6. Leiden, Brill.
Oosting, R., Dyk, J. and Glanz, O., Valence Patterns of Motion Verbs, Semantics, Syntax and Linguistic Variation, to be published.
Roorda, D. (2015a). Parallel Passages, https://shebanq.ancient-data.org/tools?goto=parallel
Roorda, D. (2015b). The Hebrew Bible as Data: Laboratory - Sharing – Experience, http://arxiv.org/abs/1501.01866
Saenz Badillos, A. (2004). *A History of the Hebrew Language*, Cambridge: Cambridge University Press.
Segarra, S., Eisen, E. and Ribeiro, A. (2013). Authorship Attribution Using Function Words Adjacency Networks, *Proc. Int. Conf. Acoustics Speech Signal Processing*: 5563-5567.
SHEBANQ, https://shebanq.ancient-data.org
Van Peursen, W. T., et al. (2015). *Hebrew Text Database ETCBC4b*. DANS. http://dx.doi.org/10.17026/dans-z6y-skyh
Young, I., Rezetko, R. and Ehrensvärd, M. (2008). *Linguistic Dating of Biblical Texts*, 2 volumes, London: Equinox Publishing.

Entities as topic labels: improving topic interpretability and evaluability combining Entity Linking and Labeled LDA

Federico Nanni
federico.nanni8@unibo.it
Data and Web Science Group, University of Mannheim

Pablo Ruiz Fabo
pablo.ruiz.fabo@ens.fr
LATTICE Lab, École Normale Supérieure, France

Introduction

Humanities scholars have experimented with the potential of different text mining techniques for exploring large corpora, from co-occurrence-based methods to sequence-labeling algorithms (e.g. Named entity recognition). LDA topic modeling (Blei et al., 2003) has become one of the most employed approaches (Meeks and Weingart, 2012). Scholars have often remarked its potential for distant reading analyses (Milligan, 2012) and have assessed its reliability by, for example, using it for examining already well-known historical facts (Au Yeung, 2011). However, researchers have observed that topic modelling results are usually difficult to interpret (Schmidt, 2012). This limits the possibilities to evaluate topic modeling outputs (Chang et al., 2009).

In order to create a corpus exploration method providing topics that are easier to interpret than standard LDA topic models, we propose combining two techniques called Entity linking and Labeled LDA; we are not aware of literature combining these two techniques in the way we describe. Our method identifies in an ontology a series of descriptive labels for each document in a corpus. Then it generates a specific topic for each label. Having a direct relation between topics and labels makes interpretation easier; using an ontology as background knowledge limits label ambiguity. As our topics are described with a limited number of clear-cut labels, they promote interpretability, and this may help quantitative evaluation.

We illustrate the potential of the approach by applying it to define the most relevant topics addressed by each party in the European Parliament's fifth term (1999-2004).

The structure of our work is as follows: We first describe the basic technologies considered. We then describe our approach combining Entity Linking and Labeled LDA. Based on the European Parliament corpus (Koehn, 2005),[1] we show how the results of the combined approach are easier to interpret or evaluate than results for Standard LDA.

Basic technologies

Entity Linking

Entity linking (Rao et al., 2013) tags textual mentions with an entity from a knowledge base like DBpedia (Auer et al., 2007). Mentions can be ambiguous, and the challenge is to choose the entity that most closely reflects the sense of the mention in context. For instance, in the expression Clinton Sanders debate, Clinton is more likely to refer to DBpedia entity Hillary_Clinton than to Bill_Clinton. However, in the expression Clinton vs. Bush debate, the mention Clinton is more likely to refer to Bill_Clinton. An entity linking tool is able to disambiguate mentions taking into account their context, among other factors.

LDA Topic Modeling

Topic modeling is arguably one of most popular text mining techniques in digital humanities (Brauer and Fridlund, 2013). It addresses a common research need, as it can identify the most important topics in a collection of documents, and how these topics are distributed across the documents in the collection. The method's unsupervised nature makes it attractive for large corpora.

However, topic modeling does not always yield satisfactory results. The topics obtained are usually difficult to interpret (Schmidt, 2012, among others). Each topic is presented as a list of words. It generally depends on the intuitions of the researcher how to interpret these tokens

in order to propose concepts or issues that these lists of words represent.

Labeled LDA

An extension of LDA topic model is Labeled LDA (Ramage et al., 2009). If each document in a corpus is described by a set of tags (e.g. a newspaper archive with articles tagged for areas like "economics", "foreign policy", etc.), Labeled LDA will identify the relation between LDA topics, documents and tags, and the output will consist of a list of labeled topics.

Our approach

Labeled LDA has shown its potential for fine grained topic modeling (e.g. Zirn and Stuckenschmidt, 2014). The method requires a corpus where documents are annotated with tags describing their content. Several methods can be applied to automatically generating tags, e.g. keyphrase-extraction (Kim et al., 2010). Our source for tags is Entity linking. Since entity linking provides a unique label for sets of topically-related expressions across a corpus' documents, it can help researchers get an overview of different concepts present in the corpus, even if the concepts are conveyed by different expressions in different documents.

Our first step is identifying potential topic labels via entity linking. Linked entities were obtained with DBpedia Spotlight (Mendes et al., 2011). Spotlight disambiguates against DBpedia, outputting a confidence value for each annotation.[2] Annotations whose confidence was below 0.1 were filtered out. We also removed too general or too frequent entities (e.g. Country or European_Union)

We then rank entities' relevance per document with tf-idf (Jones, 1972), which promotes entities that are salient in a specific subset of corpus documents rather than frequent overall in the corpus. Finally, we select the top five entities per document as per tf-idf. These five entities are used as labels to identify, with Labeled LDA, the distribution of labeled topics in the corpus.

Experiments and Results

Using the Stanford Topic Modeling Toolbox,[3] we performed both Standard LDA (k=300) and Labeled LDA (with 5 labels)[4] on speech transcripts for the 125 parties at the European Parliament (1999-2004 session). The corpus contains 125 documents, representing one party each. Documents were tokenized and lemmatised; stopwords were removed. DBpedia entities were detected with Spotlight and ranked by tf-idf, as described above.

We present the outputs of Labeled LDA with entity labels (EL_LDA) for three parties, compared to both Standard LDA and to the top-ranked entities for each party (by tf-idf). In each case, we show topics with relevance above 10%. Results for the remaining parties are available online.[5]

Only Entities - TFIDF ranked	Standard LDA	EL_LDA
Developing country Consumer Genetically modified org. Development aid Biodiversity	20%, "political term development case economic community level amendment citizen possible public question market order doe national matter regard situation" 20%, "gentleman order development lady human greens freedom food asylum citizen fundamental transport directive environment programme resource respect nuclear democracy disaster" 15%, "economic sustainable developing environmental energy local fishing investment farmer research water production consumer particularly farming oil fishery condition development agriculture" 10%, "environment amendment public agreement ensure human health directive product safety want long citizen information programme waste vote consumer industry law"	Consumer, 47% Genetically modified organism, 34% Development aid 14%

Figure: Linked entities (tf-idf-ranked), standard LDA topics and EL-LDA topics for speeches by Les Verts (France).

Entities - TFIDF ranked	Standard LDA	EL_LDA
United Kingdom Conservatism Industry Business British People	31%: "house, british, want, colleague, amendment, market, industry, united, know, business, going, hope, government, come, rapporteur, said, kingdom" 14%: ""government, ensure, economic, welcome, world, political, believe, future, common, market, directive, health, consumer, want, million, development, public, decision, farmer, food" 12%: "economic, social, public, market, measure, situation, financial, level, national, given, service, order, doe, term, community, mean, rapporteur, decision, increase, particularly"	Industry: 35% Business: 34% United Kingdom: 25%

Figure: Linked entities (tf-idf-ranked), standard LDA topics and EL-LDA topics for speeches by the Conservative Party (UK).

Only Entities - TFIDF ranked	Standard LDA	EL LDA
Basque Country Basque people Spain Nationalism Terrorism	100%, "glossed persecute inquisition underscored ulla universe exasperated unquestionable amass ddt condoned estoril cannes deceptive reappearance predominates reclassify corrects hauled remotest"	Basque People, 100%

Figure: Linked entities (tf-idf-ranked), standard LDA topics and EL-LDA topics for speeches by Partido Nacionalista Vasco (Spain).

Discussion

Labeled LDA combines the strengths of Entity Linking and standard LDA. Entity Linking provides clear labels,

but no notion of the proportion of the document that is related to the entity. Standard LDA's relevance scores do provide an estimate to what an extent the topic is relevant for the document, but the topics are not expressed with clear labels. Labeled LDA provides both clear labels, and a quantification of the extent to which the label covers the document's content.

An advantage of Labeled LDA over Standard LDA is topic interpretability. Consider the UK Conservative Party's topics. In each standard LDA topic, there are words related to the concepts of *Industry* and *Business* in general, and some words related to the UK appear on the first topic. However, in each topic, some other words (e.g. *government, directive, decision, measure, health, consumer*) are related to other concepts, like perhaps *Legislation* or *Social policy*. A researcher trying to understand the standard LDA topics is faced with choosing which lexical areas are most representative of each topic: is it the ones related to *Industry*, *Business*, and the UK, or is it the other ones? The clear-cut labels from Labeled LDA are more interpretable than a collection of words representing a topic.

The Labeled LDA topics may be more or less correct, just like Standard LDA topics. But we find it easier to evaluate a topic via questions like "is this document about *Industry*, *Business* and *the UK*, in the proportions indicated by our outputs?" than via questions like "is this document about issues like *house, british, amendment, market, industry, government,* (and so on for the remaining topics)"?

The topics for French party Les Verts illustrate Labeled LDA's strengths further. Most of the Standard LDA topics contain some words indicative of the party's concerns (e.g. *environment* or *development*). However, it is not easy to point out which specific issues the party addresses. In Labeled LDA, concrete issues come out, like *Genetically modified organism*.

Topic label *Development aid* shows a challenge with entity linking as a source of labels. Occurrences of the word *development* have been disambiguated towards the entity *Development_aid*, whereas the correct entity is likely *Sustainable_development*. These errors do not undermine the method's usefulness. Efficient ways to filter out such errors exist; this is conceptually similar to removing irrelevant words from Standard LDA topics. However, we need to be aware of and address this challenge.

Regarding Partido Nacionalista Vasco (Basque Nationalist Party), the Standard LDA topic misses the word *basque*, which is essential to this party. Labeled LDA identifies *Basque people* as a dominant concept in this party's interventions.

Outlook

Our method performs Labeled LDA using Entity Linking outputs as labels. Its main advantage is providing a specific label for each topic, that improves topic interpretability, and can simplify human evaluation of topic models.

More evaluation is needed to fully assess the approach. We will consider two possible complementary evaluations: first, a crowdsourced task where participants evaluate the coherence of Labeled LDA topics with the corpus documents. Second, an assessment of our topics by political science experts. We're mostly interested in evaluating the approach for diachronic comparisons.

Bibliography

Au Yeung, C. M. and Jatowt, A. (2011). Studying how the past is remembered: towards computational history through large scale text mining. *Proceedings of the 20th ACM international conference on Information and knowledge management.*

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. and Ives, Z. (2007). *Dbpedia: A nucleus for a web of open data*. Berlin Heidelberg: Springer.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of machine Learning research*, **3**: 993-1022.

Brauer, R., and Fridlund, M. (2013). Historicizing Topic Models, A distant reading of topic modeling texts within historical studies. *International Conference on Cultural Research in the context of "Digital Humanities"*, St. Petersburg: Russian State Herzen University.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L. and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*.

Cornolti, M., Ferragina, P. and Ciaramita, M. (2013). A framework for benchmarking entity-annotation systems. *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee.

Kim, S. N., Medelyan, O., Kan, M. Y. and Baldwin, T. (2010). Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT summit*.

Mendes, P. N., Jakob, M., García-Silva, and Bizer, C. (2011). DBpedia spotlight: shedding light on the web of documents. *Proceedings of the 7th International Conference on Semantic Systems*. ACM.

Meeks, E. and Weingart, S. B. (2013). The digital humanities contribution to topic modeling. *Journal of Digital Humanities*, **2**(1): 2-1.

Milligan, I. (2012). Mining the "Internet Graveyard": Rethinking the Historians' Toolkit.> *Journal of the Canadian Historical Association/Revue de la Société historique du Canada*, *23*(2), 21-64.

Ramage, D., Hall, D., Nallapati, R. and Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rao, D., McNamee, P. and Dredze, M. (2013). Entity linking:

Finding extracted entities in a knowledge base. *Multi-source, Multilingual Information Extraction and Summarization*. Springer Berlin Heidelberg.

Salton, G., Fox, E. A. and Wu, H. (1983). Extended Boolean information retrieval. *Communications of the ACM*, **26**(11): 1022-1036.

Schmidt, B. M. (2012). Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, **2**(1): 49-65.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, **28**(1): 11-21.

Usbeck, R., Röder, M. and Ngonga, A. C. N. (2015). Evaluating Entity Annotators Using GERBIL. *The Semantic Web: ESWC 2015 Satellite Events*. Springer International Publishing.

Zirn, C. and Stuckenschmidt, H. (2014). Multidimensional topic analysis in political texts. *Data and Knowledge Engineering*, **90**: 38-53.

Notes

[1] http://www.statmt.org/europarl/

[2] Spotlight outperforms other systems when corpus entities often correspond to common-noun mentions like *democracy*, vs. proper-noun mentions (e.g. *Greenpeace*). See Cornolti et al., 2013 and Usbeck et al., 2015.

[3] http://nlp.stanford.edu/software/tmt/tmt-0.4/

[4] Each document (party) is labeled with 5 entities. Some entities are shared across parties. For the 125 parties, this gives 300 distinct labels. This corresponds to k=300 topics in Standard LDA.

[5] https://sites.google.com/site/entitylabeledlda

Visualising Cultural Spheres – Virtual Tours and Epigraphical Data

Anna Neovesky
anna.neovesky@adwmainz.de
Academy of Sciences and Literature | Mainz, Germany

Max Grüntgens
max.gruentgens@adwmainz.de
Academy of Sciences and Literature | Mainz, Germany

Introduction

Today, cultural heritage sites, museums and other places of historical or societal value can often be visited on the Internet. Panoramic images and virtual tours allow the user to access distant sites from home via handheld devices as well as conventional desktop devices. In this way, these applications strongly reduce the threshold for getting acquainted with various cultures, their respective artefacts and unique heritage.

But can this popular and usually touristic way of presentation be used to introduce valid scientific information to a broad public? This question has been posed at the Academy of Sciences and Literature | Mainz regarding its project "Die Deutschen Inschriften".

The research project "Die Deutschen Inschriften"

The long term research project "Die Deutschen Inschriften" is a joint undertaking of six German Academies of Sciences and the Austrian Academy of Sciences. The research focuses on collecting, editing and interpreting medieval and early modern Latin and German inscriptions. They often occur in conjunction with figurative elements or spatial as well as architectural features. The inscriptions themselves are mostly in medieval Latin or in historical or regional varieties of the German language. The geographical area of research consists of Germany, Austria and South Tyrol. The inscription records range from approximately 500 AD to 1650 AD (Brandi, 1937; Kloos, 1973; Nikitsch, 2008). The project's scholars carry out their research within a wide scope of interests ranging from art history, philology and linguistics to the history of ideas. The research results are published in 90 volumes. More than 43 of these volumes, including over 17.000 records, are currently accessible through the online database "Deutsche Inschriften Online" (German Inscriptions Online, www.inschriften.net).

Virtual cultural heritage

Observing an item within a cultural heritage site in isolation frequently limits the understanding of it. This is due to its removal from the big picture of the entire ensemble in its historical, cultural and spatial context.

Two different approaches of representing historical sources in their spatial context are being explored by the projects "Inschriften im Bezugssystem des Raumes" (Inscriptions in their Spatial Context, IBR) and the virtual tours through St. Stephan in Mainz and St. Michael in Hildesheim. Project IBR utilised methods of laser scanning and semantic web technologies, in this regard aiming at a more specialised target audience. The virtual tours of St. Stephan and St. Michael on the other hand were developed as a means to visualise the spatial cultural sphere for an audience with a lower degree of specialized knowledge. In doing so the applications were generally aiming at a broader audience (Lange/Unold 2015; www.spatialhumanities.de/ibr/startseite.html; www.inschriften.net/hildesheim/ rundgang.html)

The virtual tour's objective was to arrange the scientific edition's epigraphical items in their spatial context and to put the scientific sources on display to a diverse audience in an easy accessible and comprehensible manner.