IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 2, Number 2
december 2016

Special Issue:
Digital Humanities and Computational Linguistics

Guest Editors:
John Nerbonne, Sara Tonelli

aAccademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

direttore responsabile
Michele Arnese

isbn 978-88-99982-26-3

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_2_2

IJCoL

Volume 2, Number 2
december 2016

Special Issue:
Digital Humanities and Computational Linguistics

Guest Editors:
John Nerbonne, Sara Tonelli

CONTENTS

Entities as Topic Labels: Combining Entity Linking and Labeled LDA to Improve Topic Interpretability and Evaluability

Anne Lauscher*
University of Mannheim, Germany

Federico Nanni*
University of Mannheim, Germany

Pablo Ruiz Fabo**
Ecole Normale Supérieure, France

Simone Paolo Ponzetto†
University of Mannheim, Germany

Digital humanities scholars strongly need a corpus exploration method that provides topics easier to interpret than standard LDA topic models. To move towards this goal, here we propose a combination of two techniques, called Entity Linking and Labeled LDA. 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 support the quantitative evaluation of the obtained results. We illustrate the potential of the approach by applying it to three datasets, namely the transcription of speeches from the European Parliament fifth mandate, the Enron Corpus and the Hillary Clinton Email Dataset. While some of these resources have already been adopted by the natural language processing community, they still hold a large potential for humanities scholars, part of which could be exploited in studies that will adopt the fine-grained exploration method presented in this paper.

1. Introduction

During the last decade, humanities scholars have experimented with the potential of different text mining techniques for exploring large corpora, from co-occurrence-based methods (Buzydlowski, White, and Lin 2002) to automatic keyphrase extraction (Hasan and Ng 2014; Moretti, Sprugnoli, and Tonelli 2015) and sequence-labeling algorithms, such as named entity recognition (Nadeau and Sekine 2007). The Latent Dirichlet allocation (LDA) topic model (Blei, Ng, and Jordan 2003) has become one of the most employed techniques in recent years (Meeks and Weingart 2012). Humanities scholars appreciate its capacity for detecting the presence of a set of meaningful categories called "topics" in a collection of texts (Underwood 2012; Bogdanov and Mohr 2013; Jockers 2014). Additionally, the Digital Humanities (DH) community has often remarked LDA's potential for serendipity (Alexander et al. 2014) and for distant reading analyses (Leonard 2014; Graham, Milligan, and Weingart 2016), i.e. studies that move beyond

* Data and Web Science Group - B6-26, D-68159, Mannheim, Germany. First two authors contributed equally to the paper.
** Lattice Lab (PSL Research University, USPC, ENS, CNRS, U Paris 3)
† Data and Web Science Group - B6-26, D-68159, Mannheim, Germany.
E-mail: simone@informatik.uni-mannheim.de

text exploration.
In particular, topic modeling has attracted the interest of the digital history community (Brauer and Fridlund 2013). This fascination of digital historians for a natural language processing method is a very interesting fact, as traditionally this community has focused on digital preservation, public history and geographical information systems, rather than on text analysis (Robertson 2016). We argue that this change has happened because topic modeling proposes a solution to a precise need that brings together historians as well as political scientists (Grimmer and Stewart 2013; Slapin and Proksch 2014) and other researchers whose established methodologies rely on digging in large analogue archives. Topic modeling, in its simplicity of use[1] and well hidden complexity (Underwood 2012; Weingart 2012), represents that "compass" that a historian has always needed when examining a large collection of sources. As a matter a fact, it promises to promptly offer to the researcher: *a)* a general overview of a specific collection by capturing interrelated concepts, *b)* a clear division of the collection in sub-parts (i.e. topics) and *c)* a quantification of the relevance of each topic for each document.

In the last few years, LDA has been extensively applied in digital humanities, even though it is well known that its results remain often difficult to interpret (Chang et al. 2009; Newman et al. 2010), which limits the possibilities to evaluate the quality of the topics obtained (Wallach et al. 2009). As a direct consequence of this fact, digital humanities scholars are currently stuck in a situation where they adopt topic models because they have a strong need for the potential benefits offered by such a method, especially now that large collections of primary sources are available for the first time in digital format. However, at the same time, scholars cannot derive new humanities knowledge from adopting topic models, given the current limitations of the results obtained (Schmidt 2012b; Nanni, Kümper, and Ponzetto 2016).

Specific Contribution. Given all these premises, in this paper we aim at dealing with this complex issue by providing two specific and interconnected solutions.

a) First of all, we want to provide the community with a new corpus exploration method able to produce topics that are easier to interpret than standard LDA topic models. We do so by combining two techniques called Entity linking and Labeled LDA. Our method identifies in an ontology a series of descriptive labels for each document in a corpus[2]. 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 sustain the use of the results as quantitative evidence in humanities research.[3]

b) Secondly, given the importance of assessing the quality of topic modeling results and the general lack of solid evaluation practices when adopting computational tools in digital humanities (Traub and van Ossenbruggen 2015), we provide a three-step evaluation platform that takes as an input the results of our approach and permits an extensive quantitative analysis. This will offer to digital humanities scholars an

1 See for example `http://programminghistorian.org/lessons/topic-modeling-and-mallet`.
2 We consider all Wikipedia pages as possible labels (excluding, as is usually done in entity linking, page-types like lists, disambiguation pages or redirects).
3 An implementation of the pipeline is available for download on Github: `https://github.com/anlausch/TMELPipeline`.

overview of the performance (and the drawbacks) of the different components of the pipeline, and to natural language processing researchers a series of baselines that will guide them towards improving each component of the method proposed.[4]

While the presented solutions could be adopted in various DH tasks[5], we consider the digital history community as the main target of this work. We illustrate the potential of this approach by applying it to detect the most relevant topics in three different datasets. The first dataset is the entire transcription of speeches from the European Parliament fifth mandate (1999-2004). This corpus (recently updated as Linked Open Data (van Aggelen et al. 2016)) has already been extensively adopted for computational political science research (Hoyland and Godbout 2008; Proksch and Slapin 2010; Høyland et al. 2014) and holds enormous potential for future political historians.
The second dataset is the Enron Thread Corpus, which consists of e-mail threads from the Enron Corpus, a large database of over 600,000 emails generated by 158 employees of the Enron Corporation and acquired by the Federal Energy Regulatory Commission during its investigation after the company's collapse. In the last ten years the natural language processing community has already extensively studied this dataset, conducting network and content-based analyses. Our goal is to examine the quality of our approach on a highly technical and complex dataset of a specific kind of primary source (email) that will become more and more important in future studies in the history domain.
Related to that, the third dataset is the Hillary Clinton Email Dataset, which shows a combination of features of the previous two datasets, as the Clinton emails are short correspondences mainly focused on political topics. Dan Cohen (2006), more than a decade ago, anticipated the issue that future political historians will encounter when considering the large abundance of sources[6] that public administration will leave us in the next decades. Our study intends to be a very first experimental attempt to deal with one of these new collections of primary sources, and to provide historians of the digital age with a more fine-grained text exploration solution, compared to traditional LDA.

The structure of the paper is as follows: We introduce a series of works related to our study in Section 2 and then describe our approach combining Entity Linking and Labeled LDA in Section 3. We next present the datasets (Section 4) and experiments for each component of our pipeline (Section 5). Finally, we discuss advantages and current limitations of our solution together with future work directions in Section 6.

2. Related Work

Latent Dirichlet allocation is a generative probabilistic model of a corpus, where each document is represented as a random mixture over latent topics and each topic is identified as a distribution over words. LDA can be considered an improvement of the probabilistic latent semantic analysis (Hofmann 1999) by adding the assumption that

4 The evaluation platform and the gold standard we obtained during our study is available for download on Github: `https://github.com/anlausch/TMEvaluationPlatform`.

5 We will come back later on the issues that arise when possible labels (identified in text) are missing from the knowledge base. This problem could emerge in particular when dealing with fictional characters identified in a novel, which could be not present in Wikipedia.

6 And more specifically presidential correspondences such as the 40 million email datasets from the Bill Clinton Presidential collection archived by the National Archives (Cohen 2006).

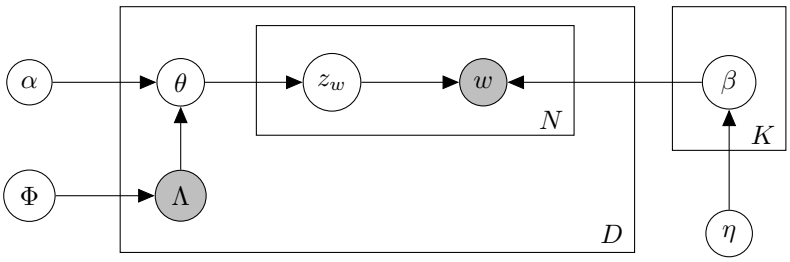

Figure 1
Plate notation of Labeled LDA.

the topic distribution has a Dirichlet prior.
During the last decade, there have been constant efforts to extend topic models by making them able to integrate or predict additional pieces of information related to the document (such as metadata information). Additionally, several studies have focused on developing methods to improve the interpretation of LDA outputs, on solutions for evaluating topic model results and on the application of topic models in humanities and social science. In the following paragraphs we will cover the solutions that are most related to our work.

Extensions of LDA. One of the first extensions of LDA is the *author-topic model* (Rosen-Zvi et al. 2004). This approach includes a specific type of metadata, i.e. authorship information, by representing each author as a multinomial distribution over topics, while each topic is associated with a multinomial distribution over words. Given a collection of documents with multiple authors, each document is modelled as a distribution over topics that is a mixture of the distributions associated with the authors. This approach was further extended to the *author-recipient-topic model* for its application in social networks (McCallum, Corrada-Emmanuel, and Wang 2005). The model not only considers individual authors, but conditions jointly on the authors of messages and on their respective recipients. Consequently, a topic distribution is assigned to each author-recipient pair.
By considering as external information the citation graph of a collection of scientific publications, the *Citation Influence Model* (Dietz, Bickel, and Scheffer 2007) is another extension of LDA that estimates the weight of edges, i.e. the strength of influence one publication has on another. The *topics over time* approach (Wang and McCallum 2006) incorporates temporal information and aims to model topic structure by identifying how this structure changes over time. Newman et al. (2006) explored the relationship between *topics and entities* (persons, locations, organisations) and introduced methods for making predictions about entities and for modeling entity-entity relationships.

Labeled LDA. The solution most closely related to our work is called Labeled LDA (Ramage et al. 2009), a supervised extension of LDA, originally used for credit attribution (namely, connecting each word in a document with the most appropriate pre-defined meta-tags and viceversa). We saw above how different types of metadata have been used in various extensions of LDA. The type of metadata exploited by labeled LDA is *keywords* associated to a document: By constraining Latent Dirichlet Allocation to define a one-to-one correspondence between LDA's latent topics and those keywords, the goal of Labeled LDA is to directly learn word-label correspondences

(taking the keywords as labels).
An illustration of Labeled LDA can be seen in Figure 1. Labeled LDA, just like standard LDA, is a generative model. It assumes that each document is generated by an imaginary process, in which both a distribution over topics for each document and, according to this, the terms in the document, are randomly selected. The plates (i.e. the rectangles) in the figure indicate replication in that generative process, whereas each node represents a random variable. The outer plate illustrates the document collection D, and the inner plate the repeated choice of topics and words for each document. The corpus-level parameters α and β are parameters of the underlying Dirichlet distribution, a multivariate probability distribution, that is the basis for LDA as well as for Labeled LDA. A word in a document is represented by a shaded node (the shading indicates that this variable can be observed). The other shaded (i.e. observable) element is the label set Λ for a document in the collection of documents D. The remaining variables are latent. In contrast to standard LDA, both the topic prior α and the label set Λ influence the topic distribution θ.
Labeled LDA has already shown its potential for fine grained topic modeling in computational social science (Zirn and Stuckenschmidt 2014). Unfortunately, the method requires a corpus where documents are annotated with tags describing their content and this meta-information is not always easily available.

Topic Labeling. Even if in the last decade the research community has strongly focused on extending LDA, exploiting various kinds of external knowledge, it has also been remarked (Chang et al. 2009) that LDA results remain very difficult to interpret for humans. Chang et al. (2009) adopted the phrase "Reading tea leaves": no better image could be found to describe how complex it can be to interpret topic model results. Given these difficulties, several researchers in natural language processing have focused on facilitating the comprehensibility of LDA results using different means, such as topic labeling[7]. One of the first papers that presents the task of topic labeling (Mei, Shen, and Zhai 2007) addresses the issue as an optimization problem involving minimizing the Kullback-Leibler divergence between word distributions, while maximizing the mutual information between a label and a topic model. A later approach (Lau et al. 2011) proposed adopting external knowledge to label topics.
The paper that is closest to our topic labeling approach (Hulpus et al. 2013) makes use of structured data from DBpedia[8]. The authors hypothesise that words co-occurring in text likely refer to concepts that belong closely together in the DBpedia graph. Using graph centrality measures, they show that they are able to identify the concepts that best represent the topics.

Entity Linking. The task of linking textual mentions to an entity[9] in a knowledge base is called *entity linking* or *entity resolution* (Rao, McNamee, and Dredze 2013). This is an information extraction task that involves being able to recognise named entities in text

7 A completely different strategy (Chaney and Blei 2012) to improve the interpretability of topic model results relies on the use of data visualisation techniques.
8 See `http://wiki.dbpedia.org/`.
9 Some authors (Chang et al. 2016) distinguish between two tasks: First, Entity Linking, where mentions corresponding to named entities are considered. Second, Wikification, where mentions to any term present in Wikipedia (even if they are common nouns) are considered. In this paper we speak of Entity Linking for both cases: We are obviating this possible difference, since it is not essential for the core idea in this work, i.e. that by tagging LDA topics with terms from a knowledge base (in this case Wikipedia), we can improve the understandability and evaluability of LDA topics.

(such as people, locations, organisations), resolving coreference between a set of named entities that could refer to the same entity (e.g. "Barack Obama" and "Mr. President") and disambiguating the entity by linking it to a specific entry in a knowledge base such as DBpedia (Bizer et al. 2009), Yago (Suchanek, Kasneci, and Weikum 2007) or Freebase (Bollacker et al. 2008). Such disambiguation process is challenging since mentions of an entity in text can be ambiguous. For this reason, entity linking systems such as TagMe! (Ferragina and Scaiella 2010), TagMe 2 (Cornolti, Ferragina, and Ciaramita 2013), DBpedia Spotlight (Mendes et al. 2011) or Babelfy (Moro, Raganato, and Navigli 2014) examine the mention in context in order to precisely disambiguate it. For instance, in the expression "Clinton Sanders debate", "Clinton" is more likely to refer to the 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*.
Since current entity linking systems rely on Wikipedia-centric knowledge bases such as, for instance, DBpedia, their performance can not always be consistent when applied to materials such as historical documents. In our work we do not focus on the entity linking step per se, and we apply it to three corpora that fit its main requirement (contemporary documents written in English). This is because we aim primarily at combining entity linking and topic models in order to improve topic interpretability. Future work will focus on expanding the potential of the entity linking step of the pipeline by permitting humanist scholars to use other knowledge bases (such as historical ontologies (Frontini, Brando, and Ganascia 2015; Tomasi et al. 2015)) to detect specific entities of interest.

Evaluation of LDA. Topic models are not simply difficult to interpret, they are also extremely complex to evaluate. Wallach et al. (2009) pointed out clearly how, even if several papers have worked on improving topic models, no single research contribution before 2009 has explicitly addressed the task of establishing measures to evaluate LDA results (Wallach et al. 2009). In order to fill this gap, they introduced two new ways of estimating the probability of held-out documents, while Mimno et al. presented a way of evaluating the coherence of the topics obtained (Mimno et al. 2011). In 2009, another highly relevant paper on the evaluation of topic models was published; this article, by Chang et al. (2009), presented the word-intrusion and the topic-intrusion tasks as evaluation practices. In these tasks, humans have to detect a word or a topic which does not fit with the rest (Chang et al. 2009). Topic models can be evaluated per se, or their results can be evaluated against a gold standard. A way of doing it is to study the alignment between topic-results and classes in the dataset, or to use topic model outputs as features in a classification task and compare it with other solutions (Nanni et al. 2016).

LDA in Humanities and Social Science. Since the end of the 1990s, and in particular after Blei et al. published their paper on LDA (Blei, Ng, and Jordan 2003), the use of LDA topic models has been widespread among the natural language processing community. Interestingly, while applying LDA to humanities and social science datasets has been attempted several times during the last decade by the NLP community (see for example Yang et al. (2011)), it is only in recent years that digital humanists and computational social scientists have started to adopt LDA as a common methodology. For instance, if we look at the proceedings of the Digital Humanities conference, we will notice that in 2011 we have the first papers applying LDA. Initial studies, such as Blevins (2010), or Jockers (2011), together with a series of introductory tutorials on topic models written by digital humanists such as Schmidt (2012), Underwood (2012) and Weingart (2012), and with a special issue of the Journal of Digital Humanities

completely dedicated to topic models (Meeks and Weingart 2012), drastically attracted the attention of the field. A similar trajectory occurred in political science. Grimmer and Stewart (2013) identified in Quinn et al. (2010) the first political science paper that adopts topic models. Afterwards, several studies such as the ones by Lucas et al. (2015), Lowe and Benoit (2013), and Zirn and Stuckenschmidt (2014) have highlighted the potential and drawbacks of the approach.

Our Contribution. Extending LDA by incorporating external information in the model and labelling topics are two tasks that have generally followed different paths. In this work, we intend to bring together beneficial elements from both tasks. Given the potential of Labeled LDA to produce topics that are easy to interpret (as they are associated with a specific label), we intend to combine it with the automatic extraction from text of entities linked to an ontology, since the potential of entities as topic labels has already been ascertained (Hulpus et al. 2013).
The drawbacks of topic models are currently limiting their adoption in humanities and social science (Trevor Owens (2012) and Nanni et al. (2016)). Taking this into account, our second contribution is to assist researchers that want to use topic modeling results as quantitative evidence by providing them with two outcomes: Not only a tool that is able to improve topic model interpretability, but also a platform that facilitates evaluating topic model quality. The platform will also be useful for researchers that intend to improve on individual components of a labeled topic modeling solution such as ours by offering a series of clearly defined baselines for each component of the pipeline.[2]

3. Method

At its heart, our approach relies on a combination of entity labels, as provided by a (possibly, off-the-shelf) entity linker, with the statistical modeling framework of Labeled LDA. As such, we do not aim to provide a new technical contribution – e.g., yet another extension of LDA – but rather to combine two well-known robust existing techniques from NLP and machine learning to provide researchers in the humanities and social sciences with easy-to-interpret statistical models of text. Basically, our methodology consists of three phases (Figure 3).

1. **Entity Linking.** Given a corpus of n documents $C = \{d_1, \ldots, d_n\}$ we initially entity-link each document to produce for each document $d_i \in C$ a set of m entity labels $E_i = \{e_{i,1}, \ldots, e_{i,m}\}$. Entity labels map arbitrary surface strings from documents into a vocabulary of entities, such as those provided by an external resource. For instance, in our setting we follow standard practices and rely on tools where the entity vocabulary is provided by a subset of the pages found in the English Wikipedia.[10] Note, however, that our method is general and can be applied with any linking system connecting words and phrases in text with an arbitrary knowledge graph.

2. **Entity Ranking and Selection.** Given a document $d_i \in C$ and its set of entity labels $E_i = \{e_{i,1}, \ldots, e_{i,m}\}$, we score elements in E_i using a weighting function $w(.)$ so as to obtain a ranked list of entities. To this end, many different weight functions can be used: in our work we rely on a simple frequency-based weighting scheme, which has been shown to be robust within standard document retrieval systems in general,

10 A *subset* in the sense that certain types of pages, e.g. lists, are not considered as possible entity candidates.

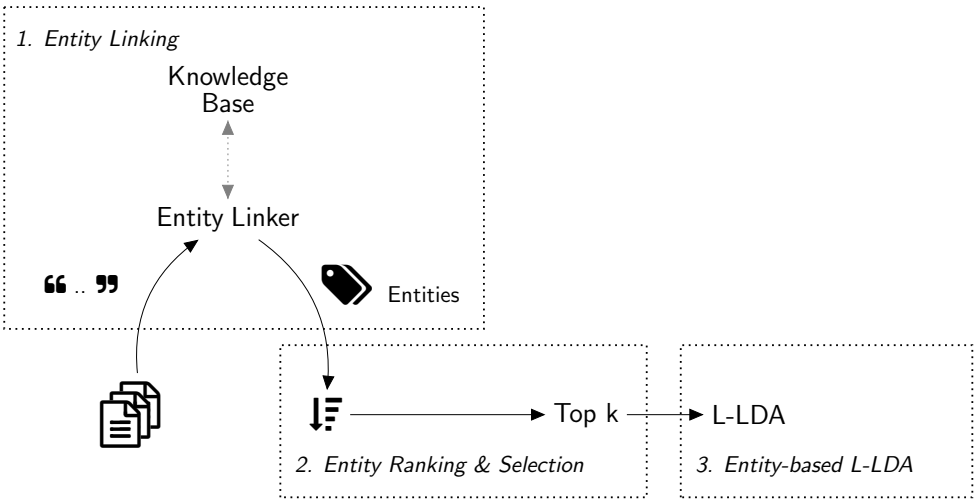

Figure 2
Schematic overview of our pipeline.

as well as query-driven entity ranking systems (Schuhmacher, Dietz, and Ponzetto 2015). This is computed as the number of occurrences of $e_{i,j}$, namely the j-th entity in the i-th document d_i, weighted over all documents in corpus C mentioning it, namely a TF-IDF weighting:

$$w(e_{i,j}) = \text{tf-idf}_{e_{i,j}} = \text{tf}_{e_{i,j}} \times \text{idf}_{e_j} = \text{tf}_{e_{i,j}} \times \log \frac{|C|}{\text{df}_{e_j}}$$

where $\text{tf}_{e_{i,j}}$ is the number of occurrences of $e_{i,j}$, namely the j-th entity in the i-th document d_i, and idf_{e_j} is the total number of documents in the corpus ($|C|$) divided over the number of documents in C that contain entity e_j (df_{e_j}). Based on TF-IDF, the weight of an entity will be highest if it has a high frequency within a few documents. Conversely, weights will be lower for those entities which occur fewer times in a document, or occur in many documents. Next, we order the entity labels in each document by decreasing score on the basis of the previously computed TF-IDF weights, and select the top k elements to provide each document with topic labels. In practice, k is a free parameter that can be set either manually (i.e., cherry-picked from the user based on data exploration) or automatically found on the basis of validation datasets or cross-validation techniques. We come back to the issue of selecting a value for k in Section 4.3 paragraph "Label Selection and Ranking".

3. **Entity-based Labeled LDA.** In the final phase of our method, we train a vanilla Labeled LDA model (Ramage et al. 2009, Section 2) using the top k entity labels from the previous step to set values in the vector of document's labels ($\lambda^{(d)}$) used to constraint the topic prior ($\alpha^{(d)}$). Given a model learned from training data, we can then apply inference techniques to test data (i.e., previously unseen documents) to compute the posterior distribution $\theta^{(d)}$ over topics per document, and accordingly rank entity labels in order of relevance for each document.

Table 1
Statistics on the corpora.

	Number of docs	Number of unique tokens	Mean number of tokens per doc	Mean number of entities per doc
EuroParl	40,192	79,332	344	21
EnronCorpus	70,178	335,032	284	8
ClintonCorpus	7,945	43,270	141	7

4. Datasets

In this section we present the three datasets we adopted in our empirical evaluation.

EuroParl. The first dataset is the entire transcription of speeches from the European Parliament (van Aggelen et al. 2016). This corpus is the result of the Talk Of Europe Project[11], which started in 2014. The goal of this project is to build a linked open dataset with data from the debates of the European Parliament, which publishes the verbatim transcripts of each speaker's statements, called 'Comptes Rendus in Extenso', as open data. Nowadays, the dataset contains all plenary debates of the fifth, sixth, and seventh terms, i.e. between July 1999 and January 2014. For the present study only the speeches from the European Parliament's fifth mandate (1999-2004) have been queried, via the SPARQL endpoint of the project[12]. Each speech was considered as a single document.[13]

EnronCorpus. The second dataset that was selected is the Enron Thread Corpus (Jamison and Gurevych 2013). This is based on the Enron dataset, a popular email dataset first presented by (Klimt and Yang 2004). However, unlike the original Enron dataset, the Enron Thread Corpus contains the emails grouped by email-thread, rather than as isolated emails. The dataset comprises over 600,000 emails generated by 158 employees of the Enron Corporation and acquired by the Federal Energy Regulatory Commission during its investigation after the company's collapse.
The email thread reconstruction on the Enron dataset, a non-trivial problem in itself, was carried out by Jamison and Gurevych, who released the corpus publicly[14]. The reason why we decided to use email threads instead of the isolated emails is that, for entity linking, precision improves when more context can be taken into account.

ClintonCorpus. The third data set is the Hillary Clinton Email Dataset. It consists of emails from Hillary Clinton's time as United States Secretary of State, during which

11 http://www.talkofeurope.eu/
12 http://linkedpolitics.ops.few.vu.nl/sparql/
13 In order to get a general overview on the collection, in a previous work (Nanni and Ruiz Fabo 2016), we represented each party as a single document, containing all speeches by its politicians. The solution chosen for this article will help us understanding the differences in topics in a more fine-grained fashion.
14 https://www.ukp.tu-darmstadt.de/data/text-similarity/email-disentanglement

she had used private email servers for official communication. The data was first published on Wikileaks in March 2016[15] and later made publicly available by the US State Department[16]. The Hillary Clinton Email Corpus represents a combination of the previous two datasets: On the one hand, it has similar characteristics to the Enron Thread Corpus, as it consists of short correspondences with limited context and highly technical content, such as for example header fields or web links. Besides, the language in both datasets is sometimes relatively informal. On the other hand, the Clinton Email Dataset contains political topics and terms, and relates therefore to the very formal speeches that can be found in the EuroParl corpus. For these reasons, the third dataset is an interesting complement to the previous two. A SQLite database containing 7,945 emails is available online, hosted by the data science platform Kaggle[17]. It consists of four tables, namely `Emails`, `Persons`, `Aliases`, and `EmailReceivers`. For the present study, the content of the `Emails` table was extracted with a custom SQL function, which exports given columns of the records of a given table to single text files, one file per entry.

By looking at the statistics on the three corpora (see Table 1), we notice that the vocabulary of the two email datasets is richer than EuroParl's vocabulary, as the email datasets are characterised by a large variety of informal expressions (e.g. "thx", "fyi", "tl:dr"). At the same time we can see that EuroParl documents are richer in terms of detected entities, given the fact that their subjects (political, economic and social concepts and actors) are largely represented in Wikipedia.

5. Quantitative Evaluation

In the next paragraphs we present the evaluation platform we designed for our empirical evaluation and the experiments we conducted for assessing the quality of our solution and for defining a baseline for future improvements.

5.1 Designing the Evaluation Platform

Document Labels. In order to assess the quality of the detected entities as labels we developed a specific browser-based evaluation platform, which permits manual annotations (Figure 3). This platform, which was initially inspired by Matsuo and Ishizuka (2004), presents a document (i.e. a political speech from the EuroParl Corpus, or an email thread from the EnronCorpus or ClintonCorpus) on the right of the screen and a set of 15 possible labels on the left (10 of them were entities present in the document and 5 of them were noisy entities from another document, all displayed in random order[18]). Annotators were asked to pick a minimum of 0 and a maximum of 5 labels that precisely describe the content of each document. In case the annotator did not select any label, this was also recorded by our evaluation system.

Entities and Topic Words. In order to establish if the selected entities were the right labels for the topics produced, we developed two additional evaluation steps. Inspired

15 `https://wikileaks.org/clinton-emails/`
16 `https://foia.state.gov/Search/Results.aspx?collection=Clinton_Email`
17 `https://www.kaggle.com/kaggle/hillary-clinton-emails`
18 Noisy entities were added in order to check whether the annotators were carefully reading the document before answering.

Figure 3
The evaluation platform we designed: Evaluating the label selection and ranking.

by the topic intrusion task (Chang et al. 2009), we designed a platform that permits to evaluate the relations between labels and topics using two evaluation modes:
For one evaluation mode (that we called *Label Mode*), the annotator was asked to choose, when possible, the correct list of topic-words given a label (see Figure 4). For the other mode (that we called *Term mode*), he/she was asked to pick the right label given a list of topic words (see Figure 5). In both cases, the annotator is shown three options: one of them is the correct match, while the other two (be they words or labels) correspond to other topics related to the same document.

5.2 Evaluation

The evaluation we performed on our solution consists of three steps, aimed at assessing the reliability of each component in our study's pipeline. First of all, we conducted a general analysis on the use of entities as document labels. Next, we evaluated the quality of our ranking. Finally, we examined the relation between topics and labels, presenting two different adaptations of the word / topic intrusion test (Chang et al. 2009), as mentioned when describing the evaluation platform in the previous section.

Label identification. Our pipeline's first step is identifying potential topic labels via Entity Linking. Linked entities were obtained with TagMe 2[19], which disambiguates against Wikipedia entities.
Thanks to the efforts of 3 human annotators, we collected 150 documents labeled with entities (50 for each dataset). The inter-annotator agreement on the label identification is $\kappa = 0.42$, which thus confirms the difficulty and subjectivity of the task[20].

19 `http://tagme.di.unipi.it` (we use standard settings – epsilon: 0.3, rho: 0.15 – for all experiments).
20 We create the final gold standard using majority vote.

Figure 4
Evaluating the Topic-Label Relation (Label Mode).

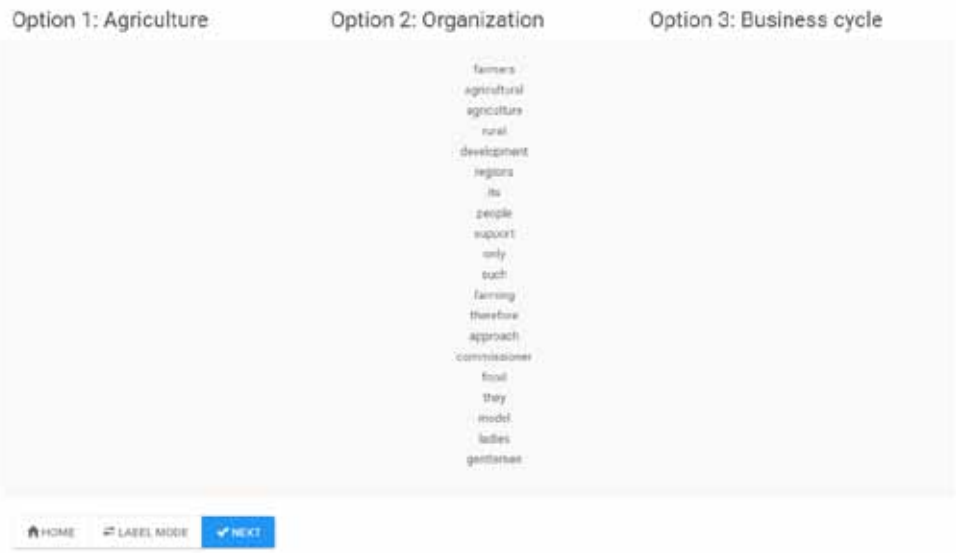

Figure 5
Evaluating the Topic-Label Relation (Term Mode).

As can be seen in Table 2, in all three datasets annotators chose four labels on average. This possibly high number of labels often occurs since two highly related entities (e.g. *Romano_Prodi* and *Politics_of_Italy*, *Employee_stock_option* and *Share_(finance)*, *Iraq* and *Iraq_War*) were both picked by the user as labels, as they reflect different aspects of the same document.
Among the labels selected, in all datasets over 90% were entities that belong to the document. It is also important to remark that selecting labels for Enron documents can be very challenging, as important aspects of the emails (like people mentioned in them)

Table 2
Evaluation of the label identification process.

	Mean number of labels selected	Precision on user selection	Number of documents without annotation
EuroParl	4.04	0.93	4
EnronCorpus	3.88	0.97	10
ClintonCorpus	4.25	0.98	1

Table 3
Evaluation of the label ranking and selection processes.

	@2			@3			@4			MAP
	P	R	F_1	P	R	F_1	P	R	F_1	
EuroParl	0.65	0.32	0.44	0.52	0.39	0.45	0.45	0.56	0.50	0.51
EnronCorpus	0.76	0.35	0.48	0.61	0.41	0.49	0.48	0.53	0.50	0.40
ClintonCorpus	0.65	0.28	0.39	0.53	0.34	0.41	0.50	0.52	0.51	0.48

are not always spotted as entities by TagMe. We will return to the impact of this issue on humanities research in the Discussion.

Label Ranking and Selection. The second evaluation step aimed at studying the quality of our label ranking and selection approach. We examined the label selection and ranking process thanks to the manual annotations obtained as described above. In particular, for establishing the quality of the label ranking we measured the Mean Average Precision (MAP) of the tf-idf ranking. Additionally, in order to assess the quality of the entity selection process, we compute precision, recall and F1-score at 2, 3 and 4 of our rankings (see Table 3)[21].
The final number of labels selected per document highly depends on the needs of the researcher. In our case, based on the results of our study (presented in Table 3) and the behaviour of the annotators (see Table 2), we select labels by considering the first 4 entities in each document, ordered by tf-idf. However, researchers with different goals (for example to obtain less topics but with high precision) can conduct the same evaluation and choose the value that best fits their needs.

Topic-Label Relation. While Labeled LDA is able to generate a specific topic for each label, it is also important to evaluate whether the label itself is representative of the

21 It is important to notice that, if we provide documents associated with only one label to Labeled LDA, its output will be the same of a Naive Bayes classifier. For this reason, we started considering the option of selecting at least 2 labels for a documents.

Table 4
Evaluation of the topic-label relation.

	Label Mode		Term Mode	
	Accuracy	# Skipped	Accuracy	# Skipped
EuroParl	0.59	2	0.38	5
EnronCorpus	0.80	4	0.75	14
ClintonCorpus	0.71	6	0.56	2

topic and vice-versa. Thanks to the effort of three human annotators, for each dataset we collected 40 annotations for the relation between labels and topic-words, 20 for each evaluation mode (i.e. choosing among topic-word sets given a label (Label Mode), or choosing among labels given one set of topic-words (Term Mode), as explained when describing the design of the evaluation platform in Section 4.2 above). In Table 4 we report the accuracy of user choices (i.e. the number of correct matches over the total number of annotated relations) and the ratio of skipped relations. As can be seen, performance differs considerably across datasets. By examining the errors we noticed that, especially in EuroParl, the user often had to decide between highly related labels (such as *Israeli-Palestinian_conflict* and *Gaza_Strip*), and this decision has a negative impact on the final results. On the other hand, while for Enron the performance on the Label Mode relation (choosing term-sets given a label) is better than for the other datasets, the annotators also skipped 14 relations in Term Mode (choosing labels given sets of topic-terms). This emphasizes how certain topics remain difficult to interpret and, thus, to associate to a label.

5.3 Discussion

The solution presented in this paper focuses on combining the strengths of Entity Linking and Labeled LDA. Entity Linking provides clear labels, but no direct notion of the proportion of the document that is related to the entity. Labeled LDA does provide an estimate to what extent the topic is relevant for the document, but it needs clear-cut labels in advance. Using entities as topic labels guarantees both clear labels, and a quantification of the extent to which the label covers the document's content.
Additionally to the quantitative evaluation presented above, we offer here a qualitative evaluation of the central steps of the process and a comparison with the results obtained using standard topic models.

Entities as Labels The decision of using entities as labels has been preferred over two other options. The first was the possibility of extracting keyphrases from text and using those as document labels. However, given the fact that we wanted to obtain clear-cut labels and to know their distribution in the corpus, we decided to take advantage of entity linking systems' capacity to identify the same entity (and therefore the same topic), even when the entity is expressed through a variety of surface mentions in the text (such as "President Obama", "Barack Obama" or "POTUS" for entity *Barack_Obama*). The second option was to consider Wikipedia categories as labels, instead of entities.

This could be done for example by labeling each document with the categories common to the majority of the detected entities. We preferred to use entities over categories, since entities offer a more fine-grained representation of the topics discussed. As an example, a combination of three entities together, such as *Barack_Obama*, *Vladimir_Putin*, and *Syria*, can easily suggest to the humanities scholar the topics of a document, without taking the risk of suggesting a too general topic by presenting *Category:Syrian_Civil_War* as the topic label.

Another aspect that was important to consider while developing the system was the reliability of entity linking systems. The quality of TagMe 2 annotations (our chosen entity linking tool) has already been studied in depth. Besides the ease of use provided by its publicly accessible web-service, the tool has been shown to reach competitive results on corpora with very different characteristics (Cornolti, Ferragina, and Ciaramita 2013), and the results are balanced whether the corpora contain mostly traditional named entities or a large proportion of concepts (i.e. terms that are found in Wikipedia but that are generally common nouns not clearly classified as a named entity). For instance, TagMe 2 achieves good results with both the IITB dataset (Kulkarni et al. 2009), where most annotations are concepts (Waitelonis, Jürges, and Sack 2016, Table 1), and with the AIDA/CoNLL dataset (Hoffart et al. 2011), where most annotations correspond to named entities (Tjong Kim Sang and De Meulder 2003, Table 2). In this work, we want to annotate concepts since we believe that conceptual information (rather than only named entities) is very useful to capture the content of our corpora (Europarl, Enron, Clinton). As TagMe 2 is good at tagging both terms for conceptual information and terms for named entities, we think that the tool is a good choice as a source of topic labels.

The most common issues we encountered when working with TagMe 2 annotations are of two types, namely wrong links and missing links. A wrong link is for example linking Gary Smith, who worked at Enron, to *Gary_Smith* the jazz-rock drummer. A missing link is not being able to identify in the ClintonCorpus "Shane Bauer", an American journalist who was detained into custody in Iran between 2009 and 2011, as he does not have a Wikipedia page.
While both wrong and missing links are errors of the system, the impact on the research conducted by the final user is very different in each case. As a matter of fact, a domain expert can easily deal with a wrong link such as *I_Need_to_Know_(Marc_Anthony_song)* by ignoring it, if it is not useful for the study, or by exploring the labeled topic, to see if it could represent a piece of meaningful information.
On the other hand, not being able to link a relevant entity because it does not appear in the knowledge base could be problematic for digital humanities scholars. While TagMe 2 performed consistently on both EuroParl and ClintonCorpus, where the majority of the detected entities are political actors and topics, this issue emerged with the EnronCorpus, where names of employees or internal resources are simply not present in a general knowledge base such as Wikipedia. For this reason, we believe it will be fundamental for the digital humanities community to focus on combining entity linking systems with domain specific knowledge bases, such as historical ontologies.

Labeled Topics The choice of using entities as topic labels provides a few advantages over simply using entity labels for a document (without topics) or just adopting standard LDA (without labels). As mentioned above, while Entity Linking provides clear labels for the content of a document, it does not offer a direct notion of the proportion of the document that is related to each entity. Conversely, standard LDA's relevance scores do provide an estimate to what extent the topic is relevant for the document, but the top-

Table 5
Linked entities (tf-idf-ranked), standard LDA topics and EL-LDA topics for speeches by the Conservative Party (UK).

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%: "industry agreement amendment situation public said government relation want health example case international concern taken product come look far" Business: 34%: "market house colleague want debate united mrs business rapporteur job hope budget government million political know view kingdom today" United_Kingdom: 25%: "amendment house government british hope citizen vote question directive welcome political national good want market debate matter law forward legal case"

ics are not expressed with clear labels. Our solution, instead, provides both clear labels, and a quantification of the extent to which the label covers the document's content. To exemplify the advantages of this solution consider the following comparison[22]. In this case we performed, using the Stanford Topic Modeling Toolbox, both Standard LDA (k=300) and Labeled LDA (with 5 labels) on speech transcripts for the 125 parties at the European Parliament (1999-2004 session)[23]. In Tables 5 and 6 we present the outputs of labeled LDA with entity labels (EL_LDA) for two 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%.

A clear 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, a researcher trying to understand the standard LDA topics is faced with choosing which lexical areas are most representative of each topic. The clear-cut labels from Labeled LDA, together with the related topic words, are more interpretable than a simple 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?" 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 the French party Les Verts, which can be seen in Table 6, 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

22 This comparison is part of the analysis we presented in our initial work on the topic (Nanni and Ruiz Fabo 2016).
23 For the 125 parties, we obtain 300 distinct labels. This corresponds to k=300 topics in Standard LDA.

Table 6
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
Developing_country Consumer Genetically_modified_organism 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"	Consumer, 47%: "public consumer political directive principle measure citizen development mean amendment national authority set economic community product respect protection know" Genetically_modified_organism, 34%: "human social health service development greens transport programme situation environment sustainable agreement democratic directive term nuclear example" Development_aid, 14%: "development human agreement trade world order community measure economic political cooperation peace life respect commitment poverty fundamental education mean essential"

easy to point out which specific issues the party addresses. In Labeled LDA concrete issues come out, like *Genetically modified organism*.
As for topic label *Development_aid*, note that it 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*. We consider that these errors do not undermine the usefulness of the overall combination approach presented here. Besides, a particularly useful aspect of an approach that combines entity labels with topics may be that topic words can provide an intuition of which aspects of a complex entity are relevant for the document. For instance, in the case of *Genetically_modified_organism*, the document addresses issues like the environment, nature and health. Going back to the case of wrong entity-labels, topic words may also perhaps provide an intuition that the entity label is not correct.

6. Conclusion

The digital humanities community has already extensively experimented with LDA topic models and has become aware of the difficulties to interpret them. In order to address this issue, we presented in this paper a combination of two techniques, called Entity Linking and Labeled LDA. 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.
In order to estimate the quality of our approach we developed an evaluation platform that permits to have a precise overview of the performance and the drawbacks of each step of our approach: label identification via Entity Linking, label ranking and selection, and the assignment of entity-labels to topics. This knowledge will help digital

humanities scholars that intend to use our solution in moving beyond text exploration studies and will offer a set of baselines to computational linguists that aim at improving each of the steps in the pipeline.
Future work on the project might include a field study, in which we would like to collect users' feedback on the system.

Acknowledgments
We want to thank all the people that used the first version of our evaluation platform at the 2016 Digital Humanities Conference in Krakow for providing us an initial gold standard and a series of useful feedbacks.
Pablo Ruiz Fabo was supported through a PhD scholarship from Région Île-de-France.

References

Alexander, Eric, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. 2014. Serendip: Topic model-driven visual exploration of text corpora. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 173–182.

Bizer, Christian, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia – A crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Blevins, Cameron. 2010. Topic modeling martha ballard's diary. *Online: http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary*.

Bogdanov, Petko and John W. Mohr. 2013. Topic models. what they are and why they matter. *Poetics*, 31:545–569.

Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Brauer, René and Mats Fridlund. 2013. Historicizing topic models, a distant reading of topic modeling texts within historical studies. In *International Conference on Cultural Research in the context of "Digital Humanities"*, St. Petersburg: Russian State Herzen University.

Buzydlowski, Jan W., Howard D. White, and Xia Lin. 2002. Term co-occurrence analysis as an interface for digital libraries. In *Visual interfaces to digital libraries*. Springer, pages 133–144.

Chaney, Allison June-Barlow and David M. Blei. 2012. Visualizing topic models. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM 2012)*, pages 419–422.

Chang, Angel X., Valentin I. Spitkovsky, Christopher D. Manning, and Eneko Agirre. 2016. Evaluating the word-expert approach for named-entity disambiguation. *arXiv preprint arXiv:1603.04767*.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, pages 288–296.

Cohen, Dan. 2006. When machines are the audience. `http://www.dancohen.org/2006/03/02/when-machines-are-the-audience/`.

Cornolti, Marco, Paolo Ferragina, and Massimiliano Ciaramita. 2013. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd international conference on World Wide Web*, pages 249–260. ACM.

Dietz, Laura, Steffen Bickel, and Tobias Scheffer. 2007. Unsupervised prediction of citation influences. In *Proceedings of the 24th International Conference on Machine Learning*, pages 233–240.

Ferragina, Paolo and Ugo Scaiella. 2010. TagMe: on-the-fly annotation of short text fragments (by Wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM.

Frontini, Francesca, Carmen Brando, and Jean-Gabriel Ganascia. 2015. Semantic web based named entity linking for digital humanities and heritage texts. In *Proceedings of the First International Workshop on the Semantic Web for Scientific Heritage at ESWC 2015*, pages 77–88.

Graham, Shawn, Ian Milligan, and Scott B. Weingart. 2016. *Exploring big historical data: The historian's macroscope*. Imperial College Press.

Grimmer, Justin and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21:267–297.
Hasan, Kazi Saidul and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1262–1273.
Hoffart, Johannes, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792.
Hofmann, Thomas. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57.
Hoyland, Bjorn and Jean-François Godbout. 2008. Lost in translation? Predicting party group affiliation from european parliament debates. *Unpublished Manuscript*.
Høyland, Bjørn, Jean-François Godbout, Emanuele Lapponi, and Erik Velldal. 2014. Predicting party affiliations from european parliament debates. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 56–60.
Hulpus, Ioana, Conor Hayes, Marcel Karnstedt, and Derek Greene. 2013. Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, pages 465–474.
Jamison, Emily K. and Iryna Gurevych. 2013. Headerless, Quoteless, but not Hopeless? Using Pairwise Email Classification to Disentangle Email Threads. In *Proceedings of 9th Conference on Recent Advances in Natural Language Processing*, pages 327–335.
Jockers, Matthew. 2011. Detecting and characterizing national style in the 19th century novel. *Digital Humanities 2011*.
Jockers, Matthew L. 2014. Topic modeling. In *Text Analysis with R for Students of Literature*. Springer, pages 135–159.
Klimt, Bryan and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European Conference on Machine Learning*, pages 217–226.
Kulkarni, Sayali, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 457–466. ACM.
Lau, Jey Han, Karl Grieser, David Newman, and Timothy Baldwin. 2011. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1536–1545.
Leonard, Peter. 2014. Mining large datasets for the humanities. *IFLA WLIC*, pages 16–22.
Lowe, Will and Kenneth Benoit. 2013. Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*.
Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2015. Computer-assisted text analysis for comparative politics. *Political Analysis*.
Matsuo, Yutaka and Mitsuru Ishizuka. 2004. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169.
McCallum, Andrew, Andrés Corrada-Emmanuel, and Xuerui Wang. 2005. The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email. *Computer Science Department Faculty Publication Series*.
Meeks, Elijah and Scott B. Weingart. 2012. The digital humanities contribution to topic modeling. *Journal of Digital Humanities*, 2(1):1–6.
Mei, Qiaozhu, Xuehua Shen, and Chengxiang Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 490–499.
Mendes, Pablo N., Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8.
Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272.

Moretti, Giovanni, Rachele Sprugnoli, and Sara Tonelli. 2015. Digging in the dirt: Extracting keyphrases from texts with kd. *CLiC it*, pages 198–203.
Moro, Andrea, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.
Nadeau, David and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.
Nanni, Federico, Laura Dietz, Stefano Faralli, Goran Glavas, and Simone Paolo Ponzetto. 2016. Capturing interdisciplinarity in academic abstracts. *To appear in D-Lib Magazine*.
Nanni, Federico, Hiram Kümper, and Simone Paolo Ponzetto. 2016. Semi-supervised textual analysis and historical research helping each other: Some thoughts and observations. *International Journal of Humanities and Arts Computing*, 10(1):63–77.
Nanni, Federico and Pablo Ruiz Fabo. 2016. Entities as topic labels: Improving topic interpretability and evaluability combining entity linking and labeled LDA. *Proceedings of Digital Humanities 2016*.
Newman, David, Chaitanya Chemudugunta, and Padhraic Smyth. 2006. Statistical entity-topic models. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 680–686.
Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *HLT: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.
Owens, Trevor. 2012. Discovery and justification are different: Notes on science-ing the humanities.
Proksch, Sven-Oliver and Jonathan B. Slapin. 2010. Position taking in european parliament speeches. *British Journal of Political Science*, 40(03):587–611.
Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.
Ramage, Daniel, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 248–256.
Rao, Delip, Paul McNamee, and Mark Dredze. 2013. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*. Springer, pages 93–115.
Robertson, Stephen. 2016. The differences between digital humanities and digital history. In *Debates in the Digital Humanities 2016*. University of Minnesota Press, pages 289–307.
Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494.
Schmidt, Benjamin. 2012a. When you have a MALLET, everything looks like a nail. *Sapping Attention*.
Schmidt, Benjamin M. 2012b. Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, 2(1):49–65.
Schuhmacher, Michael, Laura Dietz, and Simone Paolo Ponzetto. 2015. Ranking entities for web queries through text and knowledge. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pages 1461–1470.
Slapin, Jonathan B. and Sven-Oliver Proksch. 2014. Words as data: Content analysis in legislative studies. In *The Oxford Handbook of Legislative Studies*. Oxford University Press, USA, page 126.
Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, pages 697–706.
Tjong Kim Sang, Erik F. and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 142–147.
Tomasi, Francesca, Fabio Ciotti, Marilena Daquino, and Maurizio Lana. 2015. Using ontologies as a faceted browsing for heterogeneous cultural heritage collections. In *Proceedings of the 1st Workshop on Intelligent Techniques at LIbraries and Archives (IT@ LIA 2015)*, volume 5.
Traub, Myriam C. and Jacco van Ossenbruggen, editors. 2015. *Proceedings of the Workshop on Tool Criticism in the Digital Humanities*.
Underwood, Ted. 2012. Topic modeling made just simple enough. *The Stone and the Shell*, 7.

van Aggelen, Astrid, Laura Hollink, Max Kemman, Martijn Kleppe, and Henri Beunders. 2016. The debates of the European parliament as linked open data. *Semantic Web*, pages 1–10.
Waitelonis, Jörg, Henrik Jürges, and Harald Sack. 2016. Don't compare apples to oranges – Extending GERBIL for a fine grained NEL evaluation. In *Proceedings of SEMANTiCS 2016*.
Wallach, Hanna M, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112. ACM.
Wang, Xuerui and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433.
Weingart, Scott B. 2012. Topic modeling for humanists: A guided tour. *The Scottbot Irregular*, 25.
Yang, Tze-I, Andrew J. Torget, and Rada Mihalcea. 2011. Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104. Association for Computational Linguistics.
Zirn, Cäcilia and Heiner Stuckenschmidt. 2014. Multidimensional topic analysis in political texts. *Data & Knowledge Engineering*, 90:38–53.