Undefined 1 (2015) 1–5
IOS Press

Topic Modeling for Linked Open Vocabularies

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Daniel Vila-Suero [a,*], Jorge Gracia [a], Asunción Gómez-Pérez [a]
[a] *Ontology Engineering Group. Universidad Politécnica de Madrid, Spain*
E-mail: dvila,jgracia,asun@fi.upm.es

Abstract. One of the major issues still open in ontology reuse is how to help users to find the appropriate ontologies and terms for a certain application or domain of interest. In order to complement current ontology similarity-based techniques, topic modeling has the potential of allowing comparisons among ontologies not only on the basis of their lexical content, but also considering their latent semantic structure, by making emerge the topics that are implicit in their lexical descriptions.

In this paper we propose a novel method that extracts the lexical contexts of a set of ontologies, annotates them with external senses connected to the Linked Open Data cloud, and uses these annotated contexts to train a probabilistic topic model. We evaluate the method both in terms of coherence of the extracted topics and in terms of its performance when clustering topic related ontologies.

Keywords: ontology reuse, topic modeling, linked data, ontology similarity, ontology relatedness,

1. Introduction

During the last years, the increasing adoption of the linked data principles by Web practitioners has lead to a growing interconnected Web of Data, with data publishers around the world generating a web-scale data network. Among the core pillars of this emerging data network are the ontologies, or vocabularies, used for describing the data. Such vocabularies provide a set of terms to (1) describe the entities found in the data, thus (partially) describing a domain of interest, (2) make the data self-describing, and (3) help applications to integrate data from heterogeneous sources at web-scale. As stated in [18], integrating data requires bridging between the vocabularies that are used by the different data sources, and *reusing widely used vocabularies can reduce the time and effort* required for such a task. To that end, an increasing number of vocabularies for describing common things (e.g., people, places, etc.) and domains (e.g., bibliographic, geographic, or health-care) can be found on the Web. Moreover, initiatives like LOV (Linked Open Vocabularies) are a renewed effort intended to help data publishers to find vocabularies and terms, continuing the investigation started in the Semantic Web community with notable examples of ontology repositories and search engines such as Falcons [8] and Watson [11].

One of the major issues still open in *ontology reuse* is how to help users to find the appropriate terms/ontologies for a certain application or domain of interest. It has been pointed out by several works, that a key aspect for users searching and browsing ontology repositories lies in the ability of the system to show the relationships between ontologies [1], and more specifically some relationships that are implicit (i.e., not explicit relationships via imports, reuse of terms, etc.) such as their similarity to other ontologies in terms of the domain they model. However, we find that most of

*Corresponding author. E-mail: dvila@fi.upm.es

the works in the Semantic Web area dedicated to extract this type of information from ontologies, have focused either on explicit relationships or on their lexical components, using string similarity measures.

On the other hand, topic modeling techniques such as Latent Dirichlet Allocation (LDA) [5] or Biterm Topic Model (BTM) [10] have been successfully applied to extract latent topics from document collections that can be used for a number of tasks such as recommendation, document classification, and search. In particular, given a collection of documents, a probabilistic topic modeling algorithm carries out an inference process to represent documents as random mixtures over latent features, where each feature is characterized by a probability distribution over words and can be seen as a latent topic. In this way, a fixed set of K topics is represented as a probability distribution over words, and a document is represented as a probability distribution over topics.

Given the effectiveness of topic models in modeling the latent structure of large collections of documents, and the growing number of ontologies on the Web that can facilitate the application of probabilistic methods, we propose in this work a *method that applies topic modeling to ontologies, in order to reveal their latent structure*. The extraction of the ontology latent topics allows to compare ontologies in terms of topical relatedness, enabling sophisticated mechanisms to identify similar/dissimilar/complementary ontologies. Such information can be used for enhancing applications such as ontology search, ontology ranking, and ontology matching.

Moreover, our method applies word-sense disambiguation before training the topic model, with the goal of (1) reducing lexical ambiguity thus improving the quality of topics, and (2) linking the terms to the Linked Open Data (LOD) cloud, opening up new possibilities for the exploitation of topics and the improvement of the training process. In order to evaluate the feasibility of our approach, we apply our method with two state-of-the-art topic models: Latent Dirichlet Allocation [5] and the Biterm Topic Model (BTM) [10], a topic model for short texts.

Specifically, the main contributions of this paper are the following:

- We propose a novel method to perform *sense-based topic modeling of ontologies* that: (i) extracts the lexical information (context) of the ontologies, (ii) annotates such a context with senses from an external data set (BabelNet [28] in our case), and (iii) extracts, from the sense-annotated ontology contexts, the relevant latent topics and connect them to the LOD cloud.
- We provide an evaluation of our method using a state-of-the-art measure to quantify the coherence of the extracted topics for the two considered topic modeling techniques, LDA and BTM, and compare them to an LDA topic model trained with non-annotated ontology contexts.
- Additionally, we report on a task-based evaluation of the three topic models that consisted in studying the performance of the method when clustering topically related ontologies within a real corpus.

We release the open source implementation of the method and the evaluation measures, along with all the experimental data[1].

The rest of this paper is organized as follows. We start the paper by describing the related work in Section 2. Then, in Section 3, we present and describe the method. In Section 4 we illustrate the method with a real example from the library domain. In Section 5, we describe the experiments carried out to evaluate the method, and discuss their results in Section 6. Finally, we conclude the paper in Section 7.

2. Related work

In this section, we review the works related to ontology reuse. Also, the related work on ontology similarity is reviewed, as the main core technique used in ontology reuse so far. Finally, the main techniques for topic modeling are described in this section.

2.1. Ontology reuse

Recently, Schlaibe et al. [30] conducted a study to investigate the current practice of vocabulary reuse during linked data modeling. For the study, the authors identified three different aspects of the decision process when reusing vocabularies, namely: (1) *reusing vocabularies versus creating new terms and linking them to other vocabularies*, (2) *appropriate mix of vocabularies*, and (3) *usefulness of vocabulary meta-information*. Regarding meta-information, the survey results showed that the information on the number of data sets reusing the vocabulary and the information on

[1] https://github.com/dvsrepo/tovo

the domain of the vocabulary are the preferred pieces of information. In particular, a combination of both got the best results in the study. While there are already several methods to gather approximate information on popularity of a vocabulary, there is a lack of mechanisms to (semi-)automatically gather topical information of vocabularies.

In the domain of biomedical ontologies, Kamdar et al. [22] analyzed the corpus of the BioPortal repository to investigate term reuse and overlap. Their findings, consistent with the ones by Ghazvinian et al. [15], showed a high level of overlap between terms but low level of explicit term reuse and highlighted the need of more sophisticated term recommendation mechanisms that support consistent term reuse.

Alloca et al. introduced the DOOR ontology [1] to describe different types of relation among ontologies. In particular, the authors defined the *lexicographically similar to* relation which is defined as *how an ontology overlap/cover parts of the same area of interest of another ontology* by directly measuring the overlapping of terms among two ontologies. In later work [2] they hypothesized that some of the difficulties of searching ontologies are related to the information about implicit relationships between ontologies in search results. To study the impact of grouping ontology search results by types of relationships, the authors carried out a user-centered evaluation showing that from the six types of relationship studied (*comes from the same domain*, *similar to*, *is previous version of*, *is syntactically similar to* and *is included in*), *comes from the same domain* and *similar to* were the most used ontology relationships, indicating that there is a need of sophisticated mechanisms to automatically extract this type of relationships.

2.2. Ontology similarity and relatedness

In [12] David and Euzenat compared several ontology distances to calculate the similarity between ontologies. The studied metrics where either based on the structure of the ontologies or on string similarity metrics applied to their lexical elements. In later work, they proposed metrics exclusively based in the alignments between ontologies [12].

Ding et al. [14] proposed several algorithms for ranking *semantic web ontologies* to promote reuse. Their algorithms were purely based on explicit reference or links between the ontologies and no lexical information was taken into account.

Cheng and Qu [9] proposed a taxonomy for vocabulary relatedness on the Web of Data and performed an empirical study with more than 2,000 vocabularies using graph analysis techniques. Their taxonomy differentiates three kinds of relatedness: (1) *declarative relatedness*, (2) *topical relatedness*, and (3) *distributional relatedness*. In their work, *topical relatedness* measured overlap of the topics reflected in the textual descriptions of vocabularies. Textual descriptions are gathered from (1) labels of terms (e.g., rdfs:label, dc:title, and 84 inferred sub-properties) and the term IRI local name, (2) lexical descriptions of each pair of terms of the same type (i.e., a class is compared only with another class) are compared using a string metric for ontology alignment [31], The results of the analysis indicated that many topically related vocabularies are either independent copies of the same underlying conceptualization or developed by different publishers to describe a common domain.

2.3. Topic models

Researchers in the field of information retrieval have been continuously investigating ways to model information within document corpora. Topic modeling has emerged as an alternative approach to techniques such as *tf-idf* (Term Frequency - Inverse Document Frequency) with the goal of reducing the dimensionality of document descriptions and revealing the inter- and intra-document structure. One early attempt is Latent Semantic Indexing (LSI) introduced in [13]. *LSI* applies singular value decomposition to *tf-idf* descriptions of documents (i.e., a matrix composed by *tf-idf* vectors representing each document in the collection). According to the authors, LSI can capture basic linguistic features and thus reveal aspects of the generative model of the text.

Building on the idea of revealing the latent structure of document collections, in [20] the authors proposed a probabilistic variant of LSI (p-LSI), sampling the words in a document using a probabilistic model composed of multinomial random variables representing a fixed set of topics. However, p-LSI does not provide the ability of modeling the generative process within documents leading to overfitting issues and limited capability of modeling documents outside the training corpus.

Later, Latent Dirichlet Allocation (LDA) [5] has been proposed as a model to face these limitations. LDA is a generative probabilistic model where documents are represented as random mixtures over latent

features, where each feature is characterized by a probability distribution over words and can be seen as a latent topics. Since its conception, LDA and its extensions [19] have proven to be an effective and scalable solution to model the latent topics of collections of discrete data.

More recently, with the explosion of micro-blogging, Q&A services, and other social media channels, where information usually comes in the form of short and noisy text, LDA has been shown to suffer from the data sparsity of words within these kind of documents. To alleviate the data sparsity issues on short text, several alternatives have been proposed that can be classified into two categories: (1) those that make use aggregated or external information before training the topic model [32,21] (usually LDA), and (2) those that propose an alternative to the topic model itself. From the second category, a notable model is the Biterm Topic Model (BTM) [10] that has been shown to perform better than LDA, and other specific approaches for short text that assume document are drawn from one individual topic [33] seeing them as a mixture of unigrams [29].

Moreover, topic modeling has been already explored in the context of ontologies, in particular for ontology localization, as part of the work carried out in the Monnet project[2]. In Monnet, a system was developed to translate ontology labels from a given language into another [3]. The approach built on a standard *statistical machine translation* (SMT) system and investigated the effect of weighting n-grams in the language model used by the SMT system with scores obtained from topic modeling (based on LDA). This topic modeling approach was combined with :(i) the enrichment of the phrase table used by the SMT system with domain-specific translation candidates acquired from existing Web resources, and (ii) with the use of *cross-lingual explicit semantic analysis* (CLESA) as an additional technique for scoring candidate translations [4]. The experiments showed that topic modeling was not sufficient in itself to improve the SMT baseline significantly, but the combination of all the referred techniques (domain selection + CLESA + topic modeling) led to significant improvement in the translation quality.

[2] http://cordis.europa.eu/fp7/ict/language-technologies/project-monnet_en.html

3. Method

In this section we present our proposed method for training topic models for ontologies. Let us start by reviewing the main notation that we will use in the remainder of the paper. We will denote by $O = (O_1, ..., O_n)$ a given set of ontologies, and by E the set of all the ontology entities (classes, properties, and individuals) contained in such ontologies. Further, we denote as OE_i the set of all the *ontology elements* of the ontology O_i, which includes all its entities and all the elements that describe them (such as identifiers, descriptions, labels, or any other annotation). Finally, following the conventional topic modeling notation, let us define a *word w* as the smallest unit of discrete data from a vocabulary formed by {1,...,*v*}, and a *document* as a sequence of *n* words denoted by $\boldsymbol{d} = (w_1, w_2, ..., w_n)$, where w_i is the *i*th word in the sequence. We define a *corpus* as a collection of *m* documents denoted by $\boldsymbol{D} = \{d_1, d_2, ..., d_m\}$

The overall method is composed of three steps: (1) extraction of the lexical information contained in the ontologies, (2) annotation of the ontology words with external senses, (3) modeling ontology topics based on the annotated words. We detail each step in the following subsections.

3.1. Extracting lexical information from ontologies

The input to our method is a set O of ontologies. In a first step, we need to extract their lexical information and build a corpus based on it, in order to allow the later application of topic modeling algorithms.

Given the set OE_i of ontology elements of the ontology O_i, and denoting by $\mathcal{P}(OE_i)$ the power set of OE_i, we define *extraction* as a function

$$ext : E \rightarrow \mathcal{P}(OE_i)$$

that, for each ontology entity e, retrieves the descriptive information characterizing its meaning. We call *ontological context* oc_e the image of this function: $oc_e = ext(e)$.

A variety of extraction functions can be build in order to extract the ontological context of an ontology element. In our case, we extend the implementation developed in CIDER-CL [16], a tool for ontology alignment and semantic similarity computation. With this function, the extracted ontological elements comprise: URI, identifying label, synonym labels, natural language description (if available), sub-terms and super-

terms in the hierarchy, and domains, ranges and property values (if applicable). Also, a lightweight inference mechanism can be applied to extract some extra information not explicitly declared in the ontology.

In order to make such extracted ontological context useful for topic modeling algorithms, we distill the lexical information contained in the context and build a bag of words from it. Let us call $BoW(x)$ this function. We can combine the extracted lexical information from the whole ontology in this way:

$$d_i = \biguplus_{e \in O} BoW(oc_e)$$

obtaining the document d_i as result. Finally, we build the corpus $\boldsymbol{D} = \{d_1, d_2, ..., d_m\}$ based on the context extracted from every $O_i \in O$.

3.2. Annotating terms with senses

Word sense disambiguation deals with reducing lexical ambiguity by linking single-word and multiword units to their meanings [27], also known as *senses*, based on their context. The second step of our method consists in applying word-sense disambiguation (WSD) to the words contained in every ontology context $d_i \in D$, generating a set D_S of sense-annotated ontology contexts as output.

For word-sense disambiguation we use *Babelfy* [26], a tool with state-of-the-art performance that presents the additional benefit of linking senses to a large multilingual knowledge base, BabelNet [28], which integrates several wordnets, Wikidata, Wikipedia, and other large web resources. The underlying idea is that by disambiguating the words in the ontology context and by linking them to external senses, their lexical ambiguity is reduced and thus the quality of the generated topics is expected to be higher. We can use the senses to train the models and connect the generated topics to several (multilingual) knowledge bases in one step.

Although, in principle, our approach could operate with any other WSD tool, the characteristics of Babelfy make it suitable for disambiguating words coming from ontology contexts, in particular:

1. Babelfy can identify senses of words and multiwords in sequences of words with maximum length of five, which contain at least a noun. Ontology contexts extracted from classes, properties, and individuals show a significant predominance of (compound) nouns, and usually do not conform entire sentences, for instance *Person*, *Corporate Body*, or *name of person*.
2. Babelfy is a graph-based approach that relies exclusively on similarity of semantic signatures derived from the Babelnet graph, taking individual candidates from the text and not relying on the sentence structure, which can mitigate the lack of structure of ontology contexts.

The ouput of this step is a corpus D_S where each $d \in D$ is annotated with sense IDs coming from BabelNet (in BabelNet, senses are called BabelSynsets). As we use D_S to train the topic model, we filter out from each d those words that could not be annotated with BabelNet senses. Thus, finally, D_S contains sense IDs exclusively.

3.3. BTM_S, a topic model for ontologies

In this step we take the output from the WSD step D_S and use the corpus to train a topic model. The ontology contexts extracted by our extraction method have varying length and are frequently short. Moreover, after applying the WSD step these contexts become even shorter.

Taking these characteristics into account and the fact that traditional topic models such as LDA have been shown to poorly perform with corpora containing short and noisy text, we propose to use a state-of-the-art topic model, BTM, that shows a significant improvement in terms of performance when dealing with short text [10].

The effectiveness of BTM in modeling topics over short-text collections lies in the fact that it directly models word co-ocurrence patterns within the whole corpus, thus making the word co-ocurrence frequencies more stable and eventually mitigating data sparsity issues. Specifically, instead of modeling the document generation process like in LDA, BTM models the co-ocurrence of words using biterms. A biterm can be defined as an unordered combination of every two words in a document, for example given a document d with three words such as $d = (author, name, work)$, the biterm set b generated for d corresponds to $b \Rightarrow \{(author, name), (author, work), (work, name)\}$. Then, the biterm set B can be extracted for the whole corpus D by combining the biterms of each document. BTM uses the biterm set to carry out the inference of the model parameters via the collapsed Gibbs sampling method [17].

The complete description of the topic model can be found in Cheng et al., [10], but we provide details of its instantiation in Algorithm 1. In particular, in the algorithm, we introduce BTM_S, as an instantiation of BTM that combines lexical extraction from ontologies, and word-sense disambiguation. The pseudo-code describes the generative process of BTM_S using symmetric Dirichlet priors for θ and ϕ_k with α and β as hyperparameters.

Algorithm 1 BTM_S algorithm

Input: K, α, β, D_S
Output: Φ, Θ

$B \leftarrow Biterms(D_S)$
$Draw\ \Theta \sim Dirichlet(\alpha)$
for all *topic* $k \in K$ **do**
    $Draw\ \phi_k \sim Dirichlet(\beta)$
for *biterm* $b_i \in B$ **do**
    $Draw\ z_i \sim Multinomial(\theta)$
    $Draw\ w_{i,1}, w_{i,2} \sim Multinomial(\phi_i)$

4. Illustrative example

In this section, we present the application of BTM_S to a real example. Imagine a library organization that wants to publish its bibliographic catalogue as linked data. In the library domain there is a myriad of different vocabularies, ranging from those proposed by library associations such as the IFLA (International Federation of Library Associations) Functional Requirements for Bibliographic Records[3] (FRBR), or the Library of Congress BIBFRAME[4], to those created by the linked data community such as Bibo or FaBio.

Although it is indeed positive to offer several vocabularies to describe a domain, this usually leads to difficulties in deciding what vocabularies to choose and how to combine them together. One can start by searching using keywords using an ontology repository. For example, issuing two queries to the LOV repository with the keywords *edition*[5] and *workshop*[6] include Bibo[7] among the top five results. Ranking in LOV is based on a combination of popularity in the LOD cloud cloud and classical information retrieval metrics. The search results seem to indicate that Bibo is a good choice based on the keywords and its popularity (e.g., *editor* has 687,674 occurrences in 2 LOD datasets). Nevertheless, several open questions remain, are there other vocabularies describing similar entities? are there other topically related vocabularies that could complement Bibo? These kind of questions are not easy to answer within current ranked results. Let us explain our method taking this example, Bibo and another topically related vocabulary, FaBiO [8], a FRBR-based ontology. We show only samples of the outputs created by our method, the full examples can be found in the additional online material[9]

The first step (Section 3.1) is to extract the ontology context of each, that we denote as O_{Bibo} and O_{FaBiO}, and to build the following bag of words from them:

$$d_{Bibo} = (author, workshop, chapter, edition, ...)$$

$$d_{FaBiO} = (edition, preprint, created, workshop, ...)$$

The second step (Section 3.2) is to take the documents d_{Bibo} and d_{FaBiO} and to apply WSD obtaining a sense-annotated document from them. We show a sample of the sense-annotated documents in Table 1.

The final step (Section 3.3) is to take a corpus D_S and use it to train a topic model. In this example, the corpus D_S corresponds to the LOV corpus, that we will introduce in Section 5, which includes the Bibo and FaBiO ontologies. Our method will apply topic modeling to sense-annotated documents and produce the following two main outputs: *topics* and *distribution of topic within each document*

4.1. Topics

BTM_S produces a set of K topics described by their top terms t ordered by descending probability. Typical sizes to describe a topic are 5,10,20. In Section 5 we experiment with several sizes for K and top t words. In our example, if we use BTM_S to train a

[3] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records
[4] http://bibframe.org/vocab
[5] http://lov.okfn.org/dataset/lov/terms?q=edition (Last viewed 25-09-2015)
[6] http://lov.okfn.org/dataset/lov/terms?q=workshop (Last viewed 25-09-2015)
[7] Bibo, is a bibliographic ontology is described at http://bibliontology.com/
[8] http://www.essepuntato.it/lode/http://purl.org/spar/fabio
[9] https://github.com/dvsrepo/tovo

Table 1

Excerpt of Bibo and FaBiO ontology contexts annotated with the WSD method. We use the compact URI notation with the prefix bn: corresponding to the namespace http://babelnet.org/rdf/

Ontology	$Senses(terms)$			
Bibo	$bn:s00007287n$ $(author)$	$bn:s00071216n$ $(workshop)$	$bn:s00182115n$ $(chapter)$	$bn:s00029770n$ $(edition)$
FaBiO	$bn:s00029770n$ $(edition)$	$bn:s00823736n$ $(preprint)$	$bn:s00086008v$ $(created)$	$bn:s00071216n$ $(workshop)$

topic model with $K = 20$ and take a topic (i.e., topic number eight) and its top ten terms, we obtain the following list of senses and their originating words in brackets[10]:

bn:s00046705n (info)
bn:s00025314n (data)
bn:s00029345n (email)
bn:s00046516n (person)
bn:s00049910n (language)
bn:s00023236n (country,nation)
bn:s00021547n (concept)
bn:s00047172n (site,website)
bn:s00052671n (journal,periodical,magazine)

4.2. Document topic distribution

For each document in the corpus, the topic model provides its distribution by topic with the probability that the document belongs to the topic. This can be seen as a dimensionality reduction mechanism that represents the document as a mixture of topics. Using these mixtures of topics, we can calculate similarity and relatedness measures based on distances such as for example the Jensen-Shannon divergence that we introduce in Section 5.2. In our example, if we take the results of applying our method BTM_S to the D_S corpus extracted from LOV with $K = 20$, we observe that the topic distribution for Bibo and FaBiO are very similar. In particular, both present a high probability for the topic number eight, meaning that topic number eight describes well the documents, and additional probability for other topics (fourth and eighteenth), indicating other aspects of the ontology. We show below the distribution for the three most probable topics:

$$d_{Bibo} = [.., (0.1242)_4, .., (\mathbf{0.6707})_8, .., (0.0899)_{18}, ..]$$

$$d_{FaBiO} = [.., (0.0397)_4, .., (0.0308)_6, .., (\mathbf{0.8449})_8, ..]$$

In Section 4.1, we have seen that topic number eight contains senses related to personal details and publications, while topic four is similar but contains additional senses such as bn:s00044268n (history) or bn:s00056443n (music,interpretation).

4.3. Exploiting topics

Continuing with our previous example, we conclude this section by showing a direct application of BTM_S to a bigger set of library-related vocabularies. We select twenty-three vocabularies from the LOV corpus, including specialized ones such as IFLA ISBD, the RDA family of vocabularies, or BIBFRAME, as well as more general vocabularies such as schema.org, the DBpedia ontology, and the FOAF ontology.

If we train our topic model using BTM_S with $K = 20$ and the full LOV corpus, we can obtain the distribution of topics for each vocabulary, and use these distributions to analyze their differences and their relatedness. A widely-used method for measuring the similarity of two probability distributions is the Jensen-Shannon divergence (JSD) that we introduce in 5. If we apply JSD to each pair of vocabularies we can obtain a $23x23$ *divergence matrix*. Using this matrix, for each vocabulary we can obtain the most related vocabularies.

For instance, the closest vocabularies to FaBiO (ordered by ascending values for JSD) are Bibo (0.06), RDA properties for describing manifestations (0.07), and FOAF (0.09). Bibo is the closest vocabulary because it models very similar entities, especially those related to academic publishing. Regarding RDA properties for manifestation, is a FRBR-based vocabulary compose exclusively by properties to model bibliographic entities and that can in fact be used in combination with FaBiO which also FRBR-based. Also,

[10]We use the compact URI notation with the prefix bn: corresponding to the namespace http://babelnet.org/rdf/

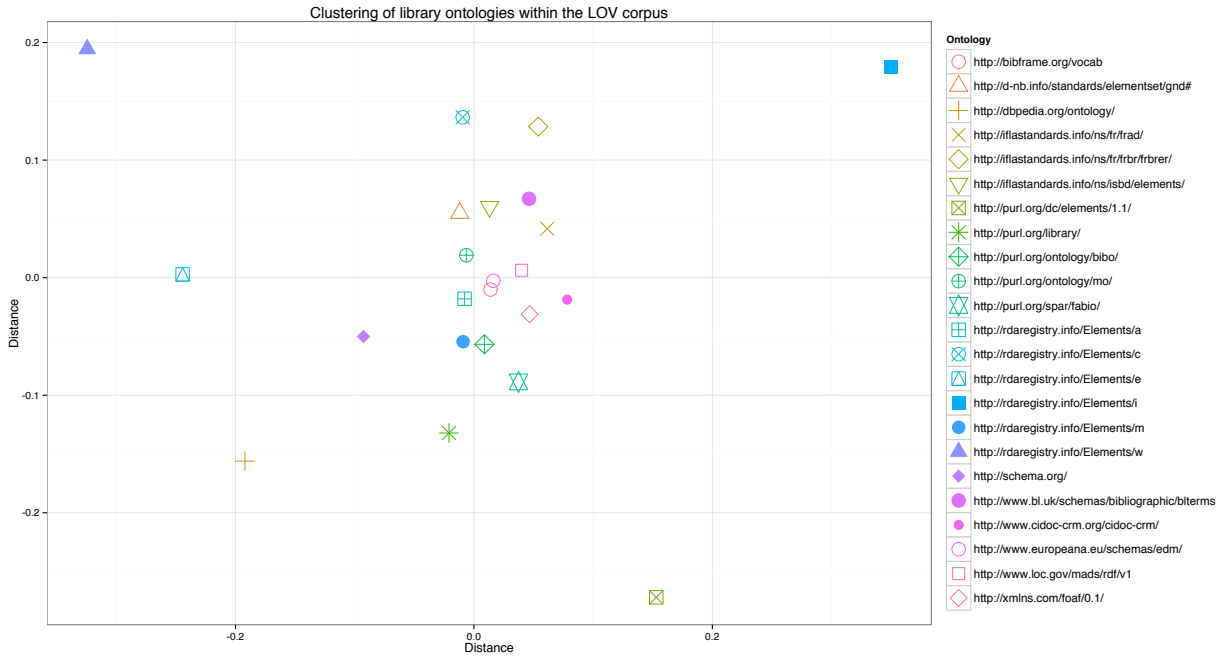

Fig. 1. An example of using topic distributions to cluster vocabularies. Two-dimensional representation of distances among vocabularies the obtained by applying multidimensional scaling (MDS) [23] to the divergence matrix based on the Jensen-Shannon divergence for every pair of vocabularies.

FOAF and FaBiO are closely related in the way they model personal information.

As we can observe, by applying standard measures such as JSD, document topic distributions can be used to calculate the relatedness between vocabularies, which can then be used to perform tasks such as ontology recommendation, or clustering.

To close this section, we show an example of ontology clustering. In order to visualize the potential clusters and verify if BTM_S is able to group together topically related vocabularies, we can apply multidimensional scaling (MDS) [23] to the aforementioned *divergence matrix*. Applying MDS with two dimensions, we obtain a graph that groups together in a two-dimensional space those vocabularies that are closer to each other. We present the results in Figure 1.

The first observation is that the majority of vocabularies clearly concentrate in a cluster, indicating that our method correctly identifies that they describe an specific domain of interest (i.e., the library domain). Another observation is that the Europeana Data Model[11] (EDM) and the BIBFRAME vocabulary are highly related (JSD=0.03). In effect, the EDM and BIBFRAME ontologies take a similar approach in terms of the entities they describe (e.g., agents, aggregations of resources, events, etc.), their granularity, and their focus on maximum coverage of bibliographic and cultural entities. It is worth noting that even when they both use different terminologies (e.g., Instance in BIBFRAME and Cultural object in EDM), BTM_S is able to identify that are highly related. This indicates that using topic distributions can provide more sophisticated mechanisms that those based on pure lexical similarity.

Also, we observe that two vocabularies are clearly distant from the others, namely the RDA properties for Works and the RDA properties for Items. By analyzing the topic distribution of the former, we observe that more prominent topic contains senses for terms such as *video*, *film*, *instrumentalist*. These terms are aligned with the purpose of the vocabulary which is to widely cover different types of creative works, as opposed to more general vocabularies to describe works such as IFLA FRBR. Regarding the latter, it is a vocabulary for describing bibliographic items or exemplars by providing physical details, which is a level of granularity barely covered by other library vocabularies.

[11] http://www.europeana.eu/schemas/edm/

5. Evaluation

In order to evaluate the strengths and weaknesses of our method, we perform several experiments, including an experiment with real-world corpora. In particular, we measure and compare three topic models:

- LDA: A traditional LDA implementation trained with non-annotated ontology contexts.
- LDA_S: The same LDA implementation[12] trained with sense-annotated ontology contexts.
- BTM_S: An implementation of the Biterm topic model trained with sense-annotated ontology contexts.

Both topic models are configured with standard values for hyperparameters α and β: $(\alpha = (50/k) + 1, \beta = 0.1 + 1)$ for LDA, LDA_S and $(\alpha = (50/k), \beta = 0.01)$ for BTM_S. The results presented in the following subsections are the result of ten runs with each method. Finally, we run the experiments with a different values for K in order to analyze the impact of the number of topics.

Although our method can work with any set of ontologies, for the evaluation we train the topic models with *Linked Open Vocabularies* corpus, a collection of more than 500 curated and topically diverse ontologies. Originally, the corpus contained 511 ontologies, but after applying the lexical extraction step we reduced their number to 504 due to parsing issues or inability of extracting lexical elements.

We have carried out two different evaluations. First, we have automatically measured the semantic coherence of the topics generated by our approach, using a state-of-the-art metric in topic modeling. Second, we have performed a task-based evaluation that consisted in using the generated topic models to perform ontology clustering. Both experiments and their results are described in detail in the following paragraphs.

5.1. Topic coherence

In order to automatically evaluate the quality of the generated topics, we apply the *topic coherence* metric proposed by Mimmo et al., [25]. This metric has the advantage of not requiring an external evaluation corpus. As we are evaluating topics extracted from ontologies that usually model varied and highly specialized domains, finding an appropriate evaluation corpus is a difficult task. Moreover, this metric has been shown to positively correlate with human coherence judgments.

The underlying idea of this measure is that the semantic coherence achieved by a topic model is related to the co-ocurrence of words describing the topics within the documents of the corpus.

Given the document frequency of word w, $f(w)$ (i.e., the number of documents with at least one occurence of word w) and the co-document frequency of word types w and w' (i.e., the number of documents containing one or more occurrences of w and at least one occurrence of w'). We can define the following topic coherence measure

$$c(t; W^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{f(w_m^{(t)}, w_l^{(t)}) + 1}{f(w_l^{(t)})}$$

where $W^{(t)} = (w_1^{(t)}, ..., w_M^{(t)})$ is a list of the M most probable words in topic t. Please note that a count of 1 is included to avoid taking the logarithm of zero.

Further, given the *topic coherence* for each topic $t \in K$, we can calculate the *average coherence* for each topic model as follows

$$C(T; K) = \frac{1}{K} \sum_{k=1}^{K} c(t_k; W^{(t_k)})$$

where T is the topic model being evaluated and K the number of topics.

In our evaluation, we calculate the *average coherence* for values of K from 10 to 60 and of M from 5 to 20. The results, presented in Figure 2, show that BTM_S consistently outperforms LDA_S for every K and length of the top M words of the topic and the improvement is statistically significant ($P-value < 0.001$).

5.2. Ontology clustering

In this section, we present the results of a task-based evaluation. In particular we would like to measure the quality of the topic models by using them to cluster topically related ontologies. The underlying idea is that if the topic model produces high-quality results, the distribution of topics for each ontology can be used to automatically group topical related ontologies. In the following paragraphs we describe the reference corpus to be used in the evaluation, and the metrics to measure the quality of the topic for ontology clustering.

[12] For LDA and LDA_S we use the implementation provided by the Apache Spark framework in its version 1.5.0

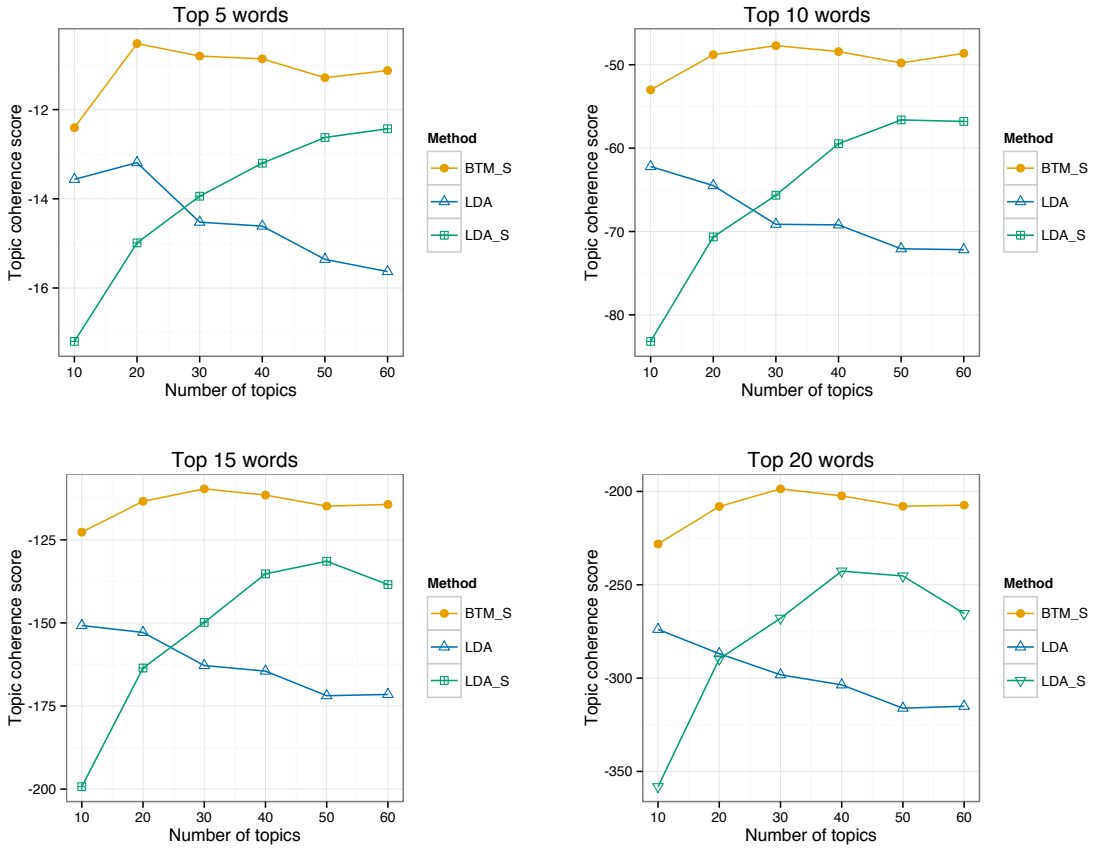

Fig. 2. Coherence scores C(T;K) results for LDA, LDA_S and BTM_S with K from 10 to 60. A larger value for the coherence score indicate more coherent topics. The average of ten runs with each method is shown.

To be able to perform this type of evaluation, we need a corpus that has been annotated with topical information for each ontology. It is worth noting that, to the best of our knowledge, there is not a corpus of ontologies directly annotated with topics or categories. Therefore, we propose a reference corpus based on the LOV corpus.

LOVTAGS is a corpus containing tags for each ontology within the LOV corpus. The tags have been extracted from the metadata dump available in the LOV website[13]. It is worth noting that these tags indicate certain aspects of the tagged ontologies and these aspects are sometimes not related to the main topic covered by the ontology. One example of topic annotation is the tag *e-commerce*, while the tag *W3C Rec* indicates the nature of the ontology publisher. Nevertheless, our goal is to evaluate the performance of our method without any intervention and we hypothesize that ontologies coming from the same ontology publisher are indeed related. Specifically, there are 42 tags with varying number of member ontologies.

Regarding the evaluation approach, there are several methods to evaluate the quality of clusters. One possible approach is to measure the distances between elements inside and outside the clusters [6].

Given that a topic model can represent a document as a probability distribution we can calculate the distance between two documents d_i and d_j, using the well-known Jensen-Shannon divergence that can be defined as follows, $dis_{JS}(d_i, d_j) =$

$$\frac{1}{2}D_{KL}(d_i||\frac{d_i+d_j}{2}) + \frac{1}{2}D_{KL}(d_j||\frac{d_i+d_j}{2})$$

[13] http://lov.okfn.org/lov.n3.gz (Retrieved 22/09/2015)

where D_{KL} is the Kullback-Leibler divergence that is defined as

$$D_{KL}(p||q) = \sum_i \log_e(\frac{p_i}{q_i})p_i$$

Further, let $G = \{G_1, ..., G_n\}$ be the set of n clusters defined in our reference corpus, given a cluster $G_n \in G$, the *intra-cluster distance* measures the distance between ontologies within a given cluster and can be defined as

$$IntraDis(G_n) = \sum_{\substack{d_i,d_j \in G_n \\ i \neq j}} \frac{2dis_{JS}(d_i,d_j)}{|G_n||G_n - 1|}$$

Applying this measure to every cluster we can evaluate the overall quality of the topic model with respect, for this we define $IntraDis(G)$ as

$$\frac{1}{N}\sum_{n=1}^{N} IntraDis(G_n)$$

On the other hand, given two clusters $G_n, G_{n'} \in G$ where $n \neq n'$, we can measure the *inter-cluster distance* that we define as

$$InterDis(G_n, G_{n'}) = \sum_{d_i \in G_n} \sum_{d_j \in G_{n'}} \frac{dis_{JS}(d_i,d_j)}{|G_n||G_{n'}|}$$

In order to evaluate the topic model with respect to the set of clusters G, we define $InterDis(G)$ as

$$\frac{1}{N(N-1)} \sum_{\substack{G_n, G_{n'} \in G \\ n \neq n'}} InterDis(G_n, G_{n'})$$

Finally, based on the notion that a quality topic model will show a smaller value of $IntraDis(G)$ value with respect to $InterDis(G)$ we define the ratio $I(G)$ as

$$I(G) = \frac{IntraDis(G)}{InterDis(G)}$$

The results of applying these measures, presented in Table 2, show that BTM_S achieves consistently better $I(G)$ for any number of topics and the improvement is statistically significant (P$-value$ < 0.001). We discuss these results in Section 6.

6. Discussion

As an overall observation, we note that the evaluation results are in line with our initial hypotheses, namely: (1) probabilistic topic modeling can be effectively applied to represent ontologies as a mixture of latent features, (2) applying word-sense disambiguation to annotate ontology contexts reduces lexical ambiguity and increases the quality of the topics, even LDA_S achieves better topic coherence than traditional LDA, and (3) BTM overcomes some of the data sparsity problems found in sense-annotated ontology contexts, achieving the best results for every experiment. Also, the usefulness of topic modeling has been highlighted throughout the paper, but its validation remains an open task for future work.

6.1. Topic coherence

Regarding topic coherence, the results highlight that BTM_S shows not only the best scores for every combination, but also that is the most stable. This stability lies in the fact that it models the generation of biterms within the whole corpus as opposed to LDA that models the document generation process. Although the results are very positive, additional evaluation has to be performed eventually involving human annotators or applying other coherence metrics, extending for example those proposed by Lau et al. [24] to make them suitable to evaluate sense-annotated topics.

Another aspect to highlight from the results is that LDA is outperformed by LDA_S starting from $K = 20$ which seems to indicate that annotating with senses can help to reduce data sparsity issues; although this would need to be verified with additional evaluations focusing on measuring this question.

6.2. LOVTAGS: task-based evaluation

Regarding the task-based evaluation, the results show that BTM_S consistently achieves the best average ratio $I(G)$ among the three evaluated techniques. Nevertheless, in absolute terms, the results could be highly improved. Therefore, to identify potential weaknesses of our approach, we perform a qualitative evaluation. In particular, we analyze the Intra-cluster distance for each cluster.

On the one hand, we observe that clusters such as *e-business*, *multimedia*, or *health* achieve very good intra-cluster scores (e.g., in the order of 0.01 in the case of *e-business* and *health*).

Table 2

Intra-cluster and Inter-cluster distances ratio $I(G)$ comparing LDA, LDA_S and BTM with the LOVTAGS corpus. The average ratio of ten runs is shown.

Corpus	Method	10	20	30	40	50	60
LOVTAGS	BTM_S	0.72±0.013	0.731±0.011	0.738±0.007	0.741±0.006	0.748±0.003	0.756±0.003
	LDA_S	0.868±0.009	0.866±0.005	0.863±0.002	0.864±0.003	0.865±0.002	0.865±0.002
	LDA	0.850±0.0053	0.851±0.0032	0.848±0.0046	0.849±0.0021	0.847±0.0032	0.846±0.0034

On the other hand, the tag *geography* achieves poor scores that are close to 1. By analyzing the data, we find that there are a number of vocabularies with a very low score that could not be correctly modeled by BTM_S due to the fact that they contain lexical information in non-English language (e.g., Norwegian, French, or Spanish). This points to the problem of multilingualism, a limitation of this work and a future line of research. The WSD method that we use can annotate senses in different languages, which make this a natural step. Besides from these vocabularies, we find that topically related geography vocabs are close to each other and in fact, BTM_S can separate in different clusters vocabularies that describe administrative units to those that describe geo-spatial features [14].

7. Conclusion and future work

In this paper we have explored the task of modeling topics within ontologies. We have proposed a novel method that extracts the lexical elements of the ontologies, annotates them with external senses, connected to the LOD cloud, and use these annotated documents to train a probabilistic topic model. Moreover, we have evaluated three topic models, traditional LDA, LDA_S and BTM_S, in order to investigate the suitability of our method for topic modeling within ontologies. Our findings reveal that BTM_S consistently outperforms the other two approaches, and produces coherent topics that can be applied to tasks such as ontology clustering, and to enhance ontology search. Besides, we release an evaluation corpus, LOVTAGS, as well as the metrics used within the paper, which can be used to evaluate further approaches.

Our future lines work include: (1) to further evaluate the topic coherence of the proposed method using human experts as proposed in [7], (2) to explore and evaluate the multilingual potential of the method in order to overcome the limitation described in the discussion section, and (3) to improve the method by leveraging the semantic structure of the sense-annotated topics. We foresee that the semantic structure of the topics can be effectively utilized to improve the performance of the training process, and for explaining and exploiting topics. As an additional line of work, we plan to implement an online version of BTM_S based on [10] to make the method more scalable and allow the modeling of unseen documents.

[14] The analysis of this cluster using a JSD divergence matrix and MDS can be found in the additional online material

Acknowledgements

This research has been funded by the LIDER FP7 project ref:610782 and the project 4V (TIN2013-46238-C4-2-R).

References

[1] C. Allocca, M. d Aquin, and E. Motta. Towards a formalization of ontology relations in the context of ontology repositories. In A. Fred, J. Dietz, K. Liu, and J. Filipe, editors, *Knowledge Discovery, Knowlege Engineering and Knowledge Management*, volume 128 of *Communications in Computer and Information Science*, pages 164–176. Springer Berlin Heidelberg, 2011.

[2] C. Allocca, M. d Aquin, and E. Motta. Impact of using relationships between ontologies to enhance the ontology search results. In E. Simperl, P. Cimiano, A. Polleres, O. Corcho, and V. Presutti, editors, *The Semantic Web: Research and Applications*, volume 7295 of *Lecture Notes in Computer Science*, pages 453–468. Springer Berlin Heidelberg, 2012.

[3] K. Asooja, T. Declerck, J. McCrae, M. Arcan, and J. Gracia. Performance evaluation and usability study v5, Monnet project (multilingual ontologies for networked knowledge). Technical report, Mar. 2013.

[4] K. Asooja, J. Gracia, N. Aggarwal, and A. Gómez-Pérez. Using Cross-Lingual explicit semantic analysis for improving ontology translation. In *Proceedings of the Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT*, pages 25–36, Mumbai, India, Dec. 2012. The COLING 2012 Organizing Committee.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[6] I. Bordino, C. Castillo, D. Donato, and A. Gionis. Query similarity by projecting the query-flow graph. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 515–522. ACM, 2010.

[7] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

[8] G. Cheng, W. Ge, and Y. Qu. Falcons: searching and browsing entities on the semantic web. In *Proceedings of the 17th international conference on World Wide Web*, pages 1101–1102. ACM, 2008.

[9] G. Cheng and Y. Qu. Relatedness between vocabularies on the web of data: A taxonomy and an empirical study. *Web Semantics: Science, Services and Agents on the World Wide Web*, 20:1–17, 2013.

[10] X. Cheng, X. Yan, Y. Lan, and J. Guo. Btm: Topic modeling over short texts. *Knowledge and Data Engineering, IEEE Transactions on*, 26(12):2928–2941, 2014.

[11] M. d'Aquin and E. Motta. Watson, more than a semantic web search engine. *Semantic Web Journal*, 2(1):55–63, 2011.

[12] J. David and J. Euzenat. Comparison between ontology distances (preliminary results). In A. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. Finin, and K. Thirunarayan, editors, *The Semantic Web - ISWC 2008*, volume 5318 of *Lecture Notes in Computer Science*, pages 245–260. Springer Berlin Heidelberg, 2008.

[13] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407, 1990.

[14] L. Ding, R. Pan, T. Finin, A. Joshi, Y. Peng, and P. Kolari. Finding and ranking knowledge on the semantic web. In Y. Gil, E. Motta, V. Benjamins, and M. Musen, editors, *The Semantic Web âĂŞ ISWC 2005*, volume 3729 of *Lecture Notes in Computer Science*, pages 156–170. Springer Berlin Heidelberg, 2005.

[15] A. Ghazvinian, N. Noy, M. Musen, et al. How orthogonal are the obo foundry ontologies. *J Biomed Semantics*, 2(Suppl 2):S2, 2011.

[16] J. Gracia and K. Asooja. Monolingual and cross-lingual ontology matching with CIDER-CL: Evaluation report for OAEI 2013. In *Proc. of 8th Ontology Matching Workshop (OM'13), at 12th International Semantic Web Conference (ISWC'13), Syndey (Australia)*, volume 1111. CEUR-WS, ISSN-1613-0073, Oct. 2013.

[17] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[18] T. Heath and C. Bizer. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136, 2011.

[19] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

[20] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[21] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.

[22] M. R. Kamdar, T. Tudorache, and M. A. Musen. Investigating term reuse and overlap in biomedical ontologies. In *Proceedings of the 6th International Conference on Biomedical Ontology, ICBO 2015, Lisbon, Portugal, July 27-30, 2015.*, 2015.

[23] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[24] J. H. Lau, D. Newman, and T. Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the Association for Computational Linguistics*, pages 530–539, 2014.

[25] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

[26] A. Moro, A. Raganato, and R. Navigli. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244, 2014.

[27] R. Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10, 2009.

[28] R. Navigli and S. P. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[29] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134, 2000.

[30] J. Schaible, T. Gottron, and A. Scherp. Survey on common strategies of vocabulary reuse in linked open data modeling. In *The Semantic Web: Trends and Challenges*, pages 457–472. Springer, 2014.

[31] G. Stoilos, G. Stamou, and S. Kollias. A string metric for ontology alignment. In *The Semantic Web–ISWC 2005*, pages 624–637. Springer, 2005.

[32] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.

[33] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer, 2011.