Evaluating the Effect of Corpus Normalisation in Topics Coherence

Luana da Silva Sousa(✉), Vinicius Melquiades de Sousa, Rogerio de Aquino Silva, and Gustavo Medeiros de Araújo

Engineering and Data Science Lab, Federal University of Santa Catarina, Florianópolis, Brazil
gustavo.araujo@ufsc.br

Abstract. Probabilistic topic models are extensively used to better understand the content of documents. Due to the fact that topic models are totally unsupervised, statistical and data driven, they may produce topics not always meaningful. This work is based on the hypothesis that, since LDA takes into account the number of occurrences of words, we could affect the quality of topics by semantically normalising the text, where each concept would be represented by the same word. We can find a formal description of lexemes found in text using a knowledgebase and extract the several forms of mentioning a lexeme to normalize a corpus. We use topic coherence metric, as it represents the semantic interpretability of the terms used to describe a particular topic, to quantify the influence of semantic corpus normalisation in topics. The first tests on the semantic normalisation framework of texts showed prominent results, and shall be investigated in depth in future.

Keywords: Corpus normalisation · LDA · Topic coherence · Ontology · Natural language processing

1 Introduction

Extracting useful information from large collections of text documents has become more challenging in recent years.

Understanding and modeling the content of documents can be very useful in many applications, such as information retrieval, natural language processing (NLP), document classification, text summarization, etc. [2].

The foundation in statistics and its capability to be extended and combined with other models make probabilistic topic model one of the most used algorithms to deal with these problems [2]. Topic modeling is a form of finding latent semantic structure within a collection of documents, and probabilistic models, such as Latent Dirichlet Allocation (LDA), have become the standard method employed [6, 20]. The intuition is that pairs of descriptor terms that co-occur frequently or are close to each other within a semantic space are likely to contribute to higher levels of coherence for a specific topic [20]. LDA model has been criticized for favoring highly frequent, general words in topic descriptors [20]. Due to the fact that topic models are totally unsupervised, statistical and data driven, they may produce topics not always meaningful [2].

E. Bisset Álvarez (Ed.): DIONE 2021, LNICST 378, pp. 197–208, 2021.
https://doi.org/10.1007/978-3-030-77417-2_15

Higher interconnectivity between information sources has the potential of increasing the utility of information. By connecting unstructured information in text documents with structured semantic data available on the internet, facts from this huge Web of Data can be used to enhance several tasks such as information extraction [25], information retrieval [27, 28], text classification [10], feature extraction [13], etc.

The goal of this work is to present the first results on a text semantic normalisation framework. Our work was based on the hypothesis that, since LDA takes into account the number of word occurrences, we could affect the quality of topics by semantically normalising the text, where each concept would be represented by the same word. If the same concept is represented by two different words in different texts, the algorithm would probably struggle more to find coherent topics. We can find a formal description of lexemes (unit of meaning, composed of one or more words) found in text using a knowledge base (KB) and extract the several forms of mentioning a lexeme to normalise our corpus. The topic coherence measure is used to address the semantic interpretability of the terms used to describe a particular topic [20], and it is the measure we used to quantify the influence of semantic corpus normalisation in topics.

1.1 Contributions

- This paper proposes a framework to semantically normalise texts, and show experimental results on topic modeling task using two widely used datasets;
- Topic coherence improvement compared to traditional LDA.

1.2 Organization of the Work

The work is organized as follows: Sect. 2 presents a background content of the methods approached in this paper. Section 3 brings related work and compares our approach to others in literature. Section 4 exposes the proposed method to semantically normalise text and how it was applied to topic modeling. Section 5 presents the results and a discussion. And finally, Sect. 6 concludes our work.

2 Methods

2.1 Topic Modeling and Topic Coherence

A topic is a probability distribution over words and documents are mixtures of topics. Hence, a topic model can be considered a generative model for documents [24]. A more formal description of the Topic Modeling problem using LDA model is described as follows.

In LDA, it is assumed that there are K underlying topics from which the documents are generated and that each topic is represented as a multinomial distribution over the V words in the vocabulary. Therefore, a document is generated by sampling a mixture of these topics and then sampling words from that mixture [6].

A document with N words $d = w_1, \ldots, w_N >$ is generated by the following process:

1. The mix of topics θ is sampled from a Dirichlet distribution $(\alpha_1, \ldots, \alpha_k)$;
2. For each of the N words, a topic $z_n \in \{1, \ldots K\}$ is sampled from a $Mult(\theta)$ distribution, where $p(z_n = 1 \vee \theta) = \theta_i$;

3. Each word w_n is sampled, conditioned to the z_n-th topic, from the multinomial distribution $p(k \vee z_n)$.

It is possible to think of θ_i as the degree that a topic refers to a document. So, the probability of a document is the following mix:

$$p(d) = \int_{\theta} \left(\prod_{n=1}^{N} \sum_{z_n=1}^{K} p(w_n|z_n; \beta) p(z_n|\theta) \right) p(\theta|; \alpha) d\theta \tag{1}$$

Where $p(\theta; \alpha)$ is Dirichlet, $p(z_n \vee \theta)$ is a multinomial distribution parametrized by θ, and $p(w_n|z_n; \beta$ is a multinomial distribution over words. This model is parametrized by the parameters $\alpha = \langle \alpha_1, \ldots, \alpha_k \rangle$ and a matrix β with dimensions $K \times |V|$. The per-word topic assignment, per-document topic distribution and topics are all latent variables and are not observed. The only observed variable is words within the documents, to infer the hidden structure (latent variables) with statistical inference [26].

As a way of making it clearer, Fig. 1 depicts the word distribution over the topics and a topic distribution over documents. As it was said before, a document is a distribution of topics and a topic is a distribution of words.

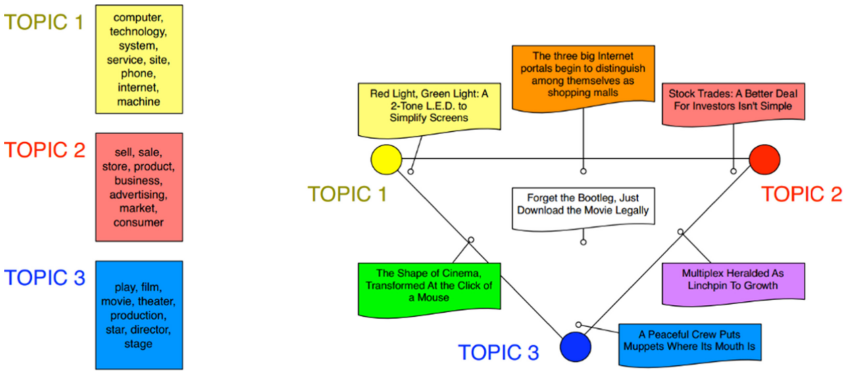

Fig. 1. Word-Topic distribution on the left and Document-Topic distribution on the right. These three topics represent the first three topics from a fifty topic LDA model trained on articles from the New York Times [8].

The topic quality (quality means interpretable and meaningful), measured as topic coherence, is based on the hypothesis that words with similar meaning tend to co-occur within a similar context. Each topic distribution contains every word but assigns a different probability to each of the words. The words with the highest probabilities within a topic are those that tend to co-occur more frequently. So, the top 10 or 15 high-probability words are usually used to interpret and semantically label the topics [26].

Researchers use several metrics of model fit, such as perplexity or held-out likelihood. However, such measures are only useful for evaluating the predictive model and do not address the explanatory goals of topic modeling [8]. The task of quantifying the coherence of a set of topics have been studied to remedy the problem that topic models give no guaranty on the interpretability of their output [23].

Many measures of coherence have been proposed recently, based on approaches that include co-occurrence frequencies of terms within a reference corpus [16, 19, 23]. A recent study [23] systematically and empirically explored the multitude of topic coherence measures and their correlation with available human topic ranking data. Their approach revealed a new coherence measure, called C_V, which achieved the highest correlation compared with all human ranking data. Hence, this study adopts the C_V coherence measure for topic coherence calculations.

2.2 Semantic Web

The idea of Semantic Web was described in 2001 by Tim Berners-Lee et al. as "A new form of web content that is meaningful to computers". In this new form of web content, introduced as an extension of the current web, information is given a well-defined meaning, where computers and people can work in cooperation [4].

The Semantic Web is based on the Resource Description Framework (RDF), a formal language for describing structured information [15]. An RDF document describes a formal specification of an arbitrary domain. This specification is modeled by a directed, labeled graph where each edge represent a link between two resources, represented by the graph nodes [17]. The link is expressed as RDF triples (*subject*, *relation*, *object*). Uniform Resource Identifiers (URI) are used to identify RDF resources and relations. To access and query RDF graphs the Protocol And RDF Query Language (SPARQL) was developed [21]. The results of SPARQL queries can be new RDF graphs or sets of resources.

The relationships and properties RDF resources may have can be specified by the vocabulary description language RDF Schema (RDFS) [7]. RDFS allows to create custom defined vocabularies to organize knowledge. Since URIs enable to identify RDF resources globally, it seems reasonable to combine vocabularies shared by different creators and across different domains. When shared, an RDF vocabulary can be denoted as an *ontology*. An ontology is an explicit, formal specification of a shared conceptualization and defines the terms used to describe and represent an area of knowledge [14].

The concept of ontology brings us to the Linking Open Data (LOD) project. It aims to identify datasets in the web that are available under open licenses, re-publish these datasets in RDF and interlink them with each other [5]. The term Linked Data refers to a set of principles to publish and interlink structured data on the web. One of the ontologies available on the web is YAGO (Yet Another Great Ontology - https://yago-knowledge.org/). YAGO is a large semantic knowledge base, derived from Wikipedia, WordNet, WikiData, GeoNames, and other data sources [22]. Currently, YAGO knows more than 17 million entities (like persons, organizations, cities, etc.) and contains more than 150 million facts about these entities. SPARQL queries are used in this work to query Yago Knowledge base in order to fetch alternative words for the same lexeme.

2.3 Named Entity Linking

Named Entity Linking can be described as the task of identifying lexemes in a text and linking them to the entity they name in a knowledge base, such as DBPedia. Before going

too deep, an introduction of terminology and concepts is established. The term *entity* refers to something which is cognitively representable. An entity *mention* refers to the part of the text where a reference to an entity is made. It is also called *lexeme*, which is the basic unit of meaning. The *surface form* is a specific syntactic representation of the lexeme (the exact character string). A *knowledge base entity* refers to a representation of the entity, usually identified by an *URI* [28].

Now, let K be a formal knowledge base, $d \in D$ a document of the corpus D, $W \subseteq d$ the words of document d, $M \subseteq 2^W$ the set of entity mentions, and $m = (s, l, d, c) \in M$ denote an entity mention in a document d with start position s, length l and confidence score $c \in [0, 1]$. The *named entity linking problem* can be described as this [28]:

Definition 1 (Name Entity linking Problem)

- An extraction function $f_{ex} : W \rightarrow M$ to extract the entity mentions M from a document set D.
- A mapping function $f_{map} : M \rightarrow 2^W \cup NIL$ to compile a list $C \in 2^K$ of potential knowledge base entity candidates for every lexeme.
- A scoring function $f_{score} : C \rightarrow R$ to calculate a score, which indicates the degree of certainty that the candidate URI is to be selected as the correct one.
- A selection function $f_{sel} : C \rightarrow K$ to select the right candidate according to the calculated scores.

The degree of ambiguity is indicated by the size of the candidate list C. Hence, the *disambiguation task* is described by putting the mapping, scoring and selection functions together. The entire *context* is observed when processing the analysis items in the implementation of these functions. Just like in communication theory and linguistics the context is essential when interpreting pieces of information, in NEL it is as well. Examining context is crucial for NEL, because some context items can be very decisive when interpreting the context information [28].

There are some options of automated entity linking, and one of them is DBpedia Spolight (https://www.dbpedia-spotlight.org/) [18]. It is and open source project developing a system for automatic annotation of DBpedia entities in natural language text. It provides an interface for phrase spotting (recognition of phrases to be annotated) and disambiguation (entity linking) as well as various output formats (XML, JSON, RDF, etc.) in a REST-based web service [9]. DBpedia Spotlight is used in this work as the tool to find resources (URIs corresponding to formal descriptions of a concept) in text. These resources have several information and metadata about the concept, as well as links to other knowledge bases.

3 Related Work

There are related works of a type of topic modeling called of knowledge-based topic modeling. The main difference with the known knowledge-based topic modeling [3, 12] is that in this work the knowledge-based content is nor on the sampling neither on the inference steps [11, 29], it is a preprocessing step, applied to the input text.

Furthermore, there have been lots of works trying to solve different NLP tasks using semantics [18]. Short text classification was dealt by [12]. They exposed the use of DBpedia ontology to better represent short texts, so that semantically similar texts with no words in common can have similar context [12]. Their approach consisted in three steps: (i) identify concepts in text using DBpedia Spotlight and annotate them as resources; (ii) select the concepts with higher similarity; and (iii) extract additional knowledge, like categories, types or topics of identified concepts. The main dissimilarity between [12] and this word is that they added the additional knowledge (additional words) to the text and did not normalise the text. Moreover, they tested their hypothesis in a classification task, and not in a topic modeling context.

[29] proposed a knowledge-based topic modeling based on multi-relational knowledge graphs. They proposed a method that models document-level word co-occurrence with knowledge encoded by entity vectors automatically learned from external knowledge graphs. In other words, they do not consider only lexemes recognized in text, but from triples in external knowledge graphs. Our work is different from [29] because they add semantic knowledge into the generative process and not in preprocessing. Yet to the best of our knowledge, this is the first work that semantically normalise documents using a semantic replacement methodology.

4 The Proposed Method

There are two big steps that compose this method: (i) semantic corpus normalisation and (ii) topic modeling. The first one is the main contribution of this work, where we semantically normalised a corpus in order to benefit from the explicit semantics of Linked Data to evaluate the effect on the coherence of topics; while the second is the method used to show how semantically normalised texts affect semantic coherence of NLP tasks such as topic modeling using small texts.

4.1 Corpus Normalisation

Figure 2 depicts the normalisation method architecture, where each box is explained in the following paragraphs.

The normalisation is composed of two steps: (i) Resource extraction and (ii) Transformer. This first step is to find all lexemes in texts and associate them with resources from DBpedia. Lexemes are annotated by the process of NEL, using the DBPedia Spotlight annotation tool. Once we have the possible resources mentioned in the text and its respective URIs, we can find additional knowledge related to this resource. We decided to search in another knowledge base in order to find potential alternative surface forms, such as other labels used to describe that resource. The Yago KB is used in this step as the authors found more options in *alternate Name* and *label* fields in this KB. The second step is to create a replacement data structure, where all possible labels of a resource would be replaced by only one. Finally, the last step is to replace them all.

Since the resource extraction is achieved by making HTTP requests and SPARQL queries for each document separately, it is modularized and convenient to parallelize. The documents are saved in a database, each one with an associated unique identifier. We

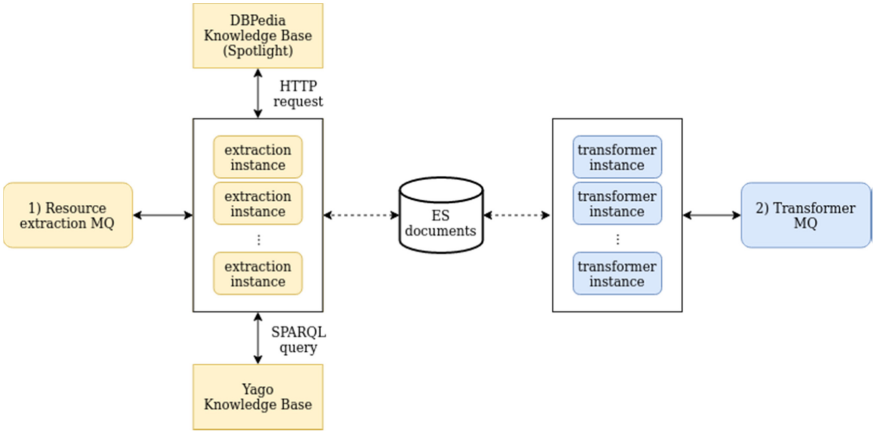

Fig. 2. Normalisation Architecture. The yellow blocks refer to the resource extraction step, as the blue block refers to the transformation step.

used Elasticsearch (https://www.elastic.co/) [1] as database. This identifier is used to keep track of which documents had already been processed. The RabbitMQ (https://www.rabbitmq.com/) tool is used to coordinate the extraction, creating a queue of documents to be processed. The resource extraction module runs in several instances (left side of Fig. 2), in order to accelerate the extraction. Each instance consumes from the queue, represented in the Fig. 2 as *Resource extraction MQ* in order to know which text it should process next.

The extraction works in this way for each text: first, it annotates all resources found in text using DBpedia Spotlight; second, for each resource, it makes a SPARQL query to Yago Knowledge base searching for alternative labels and the labels registered for that resource; lastly, it aggregates all possible labels for a resource and save them in the database.

The transformer step, which is done once all resource extraction is over, collects all resources and labels and organize them in a big mapping list. The mapping list maps all possible labels of a concept to a main label, which is going to replace all possible mentions of that concept. Once the mapping list is built, a regex substitution task is performed in order to make all substitutions. A queue is used to manage all texts that are being processed, similar to the resource extraction phase.

4.2 Topic Modeling

The first step to extract topics is the preprocessing one. The following preprocessing is done: (i) remove invalid characteries and punctuation; (ii) lowercase; (iii) tokenize (transform text into a word vector); (iv) remove stopwords (too common words that do not aggregate meaning); (v) form bigrams (composed words, e.g. “United States”) and (vi) lemmatize (remove word inflections, returning it to its root form, e.g. “said” to “say”). Besides usual stopwords, a list of too frequent words is removed too. From 20-Newsgroup: from, subject, re, edu, use, not, would, say, could, _, be, know, good,

go, get, do, done, try, many, some, nice, thank, think, see, rather, easy, easily, lot, lack, make, want, seem, run, need, even, right, line, even, also, may, take, come; and from Reuters: from, subject, re, edu, use, say, inc, -PRON-.

After preprocessing, the vocabulary of words is ready to compose the word-document matrix that serves as input to LDA algorithm.

5 Results and Discussion

We used two very known corpus of NLP tasks: 20-Newsgroups (https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html) and Reuters (https://www.nltk.org/book/ch02.html). The 20-Newsgroups has more than 18.000 newsgroups posts on 20 topics. Its is divided in training and testing, although for this work we used both as an unique dataset. As the news were from 20 topics, we also used 20 for the hyper-parameter of topics. The Reuters Corpus contains more than 10.000 news documents totaling 1.3 million words. The documents have been classified into 90 topics, and grouped into two sets, called "training" and "test". However, for this work we use both training and test set to extract the topics. Also, as the original corpus was annotated in 90 topics, we used 90 for the hyper-parameter of topics.

In Fig. 3 we can see the distribution of words per document in each corpus used. There is a bigger variety of sizes in 20-Newsgroup corpus, as well as the document length mean is higher before preprocessing. After preprocessing the remaining useful words were similar between both corpora.

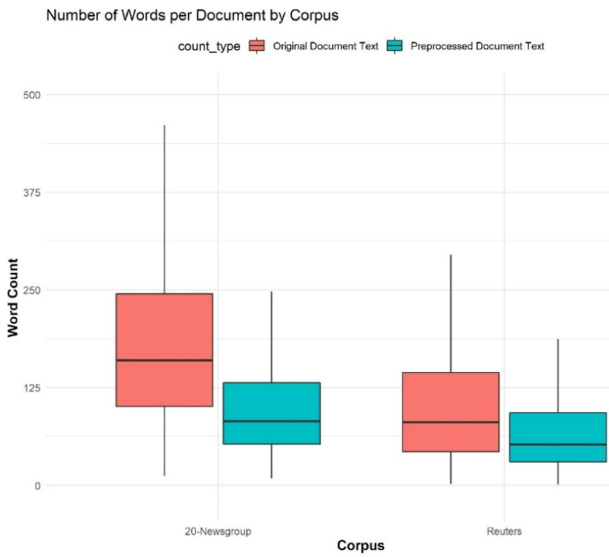

Fig. 3. Number of words per document by corpus. The red boxplot shows the counter of all words, separated only by spaces. The blue one shows the preprocessed documents, where stopwords and bigrams were built. This preprocessing is the same the documents are exposed before topic modeling algorithm.

In Table 1 there are examples of both corpora. The 20-Newsgroup has a form of e-mails, short texts and Reuters has a form of article documents. It can be seen on Fig. 3 that after preprocessing, 20-Newsgroup has lost more words than Reuters, because a big amount of characters were not letters or digits, which are removed on the preprocessing step.

Table 1. Document examples of each corpus.

20-Newsgroup	Reuters
"From: Edwin Gans Subject: Atheism Nntp-Posting-Host: 47.107.76.97 Organization: Bell-Northern Research Lines: 1"	"AMERICAN CENTURY \< ACT > RESTATES EARNINGS\n American Century Corp said it\n has restated its earnings for the fiscal year ended June 30,\n 1986 to provide an additional five mln dlrs to its loan loss\n allowance, causing a restated year-end net loss of 14,937,000\n dlrs, instead of 9,937,000 dlrs.\n The company said the change came after talks with the\n Securities and Exchange Commission on the company's judgement\n in considering the five mln dlrs collectible.\n In the note to its 1986 financial statement, American\n Century said it considered the five mln dlrs collectible,\n making its loan loss provision less than required.\n The company said in spite of the SEC decision, it still\n feels its allowance for possible loan losses at June 30, 1986\n was adequate and that it has considered all relevant\n information to determine the collectability of the five mln dlr\n receivable.\n But, it said continued disagreement with the SEC staff\n would not be in its best interest.

After a minimal analysis of the corpora used, the resources and possible labels were extracted from text and saved into the database. With all possible labels saved, the mapping list was built and used to transform the texts. The results for this experiment are shown on Table 2. The topic coherence for 20-Newsgroup corpus decreased with the corpus normalisation, as for Reuters corpus the coherence increased from 0.456 to 0.475.

Table 2. Topic coherence of the top 10 words in topic using C_V measure.

Dataset	Original corpus	Normalised corpus
Reuters	0.456	0.475
20-Newsgroup	0.672	0.667

As it can be seen by the results in Table 1, there is a positive effect on topic coherence on Reuters corpus, while on 20-Newsgroup it seems to have decreased the metric. From this results, we can leverage a number of hypothesis for these differences: (i) the size of the documents matters, because in small texts it is more difficult to get resources due to the fact that there is little context for the algorithm to disambiguate resources; (ii) the nature of text, as 20-Newsgroup has an e-mail like writing and Reuters is more article like; or (iii) the completeness of the knowledge base in specific topics. On the first hypothesis on the size of documents we can say that when a document it too small, the algorithm cannot be confident enough that a lexeme corresponds to a resource, so it does not capture it. Although 20-Newsgroup has a higher length of documents, both corpora are small, with a mean of less than 200 words per document. Also, by the Fig. 3 we can see that the number of valid words decrease much more on 20-Newsgroup than on Reuters corpus. Hence, we can infer that, although the total number of words is bigger on 20-Newsgroup, the number of valid lexemes to the algorithm to extract resources is very close to Reuters corpus. Besides that, as the context matters, and the context is the set of words around a lexeme, it is very difficult to the NEL algorithm to link a useful resource to the lexemes in text if only just a few words are valid.

This leads to the second hypothesis on the nature of text. It can be seen by Table 1 that the texts have very different natures. An e-mail like text is much more prone to have symbols and initials or acronyms, as seen in the first text of 20-Newsgroup of Table 1. On Reuters, it can be seen that the text is more fluent and without many symbols.

Related to the third hypothesis, we can explore in the future implementations a more complete log to track the resources that exist in the KB or not. The authors noticed during the execution of tests that many resources from Yago linked in the DBpedia page were not available anymore.

6 Conclusion

In this work we presented a framework to semantically normalise texts using resources from the Semantic Web. Our framework was tested in a topic modeling problem using two known corpora in order to have the first results and take insights for improvements.

The framework for normalisation is capable of improving the topic coherence of one of the corpora being tested.

So, the first tests on the semantic normalisation framework of texts showed prominent results and shall be investigated in depth in future. The authors plan to test this normalisation framework on a larger corpus from scientific articles or Wikipedia pages, in order to improve the analysis on the first and second hypothesis.

References

1. ELASTICSEARCH (2019). https://www.elastic.co/pt/
2. Allahyari, M.: Semantic Web Topic Models: Integrating Ontological Knowledge and Probabilistic Topic Models. Ph.D. thesis, University of Georgia (2016)
3. Allahyari, M., Kochut, K.: Semantic tagging using topic models exploiting wikipedia category network. In: 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), pp. 63–70. IEEE (2016)

4. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Sci. Am. **284**(5), 28–37 (2001)
5. Bizer, C., Heath, T., Idehen, K., Berners-Lee, T.: Linked data on the web(ldow2008). In: Proceedings of the 17th International Conference on World WideWeb, pp. 1265–1266 (2008)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
7. Brickley, D., Guha, R.V., McBride, B.: RDF schema 1.1. W3C Recomm. **25**, 2004–2014 (2014)
8. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Advances in Neural Information Processing Systems, pp. 288–296 (2009)
9. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems, pp. 121–124 (2013)
10. De Melo, G., Siersdorfer, S.: Multilingual text classification using ontologies. In: European Conference on Information Retrieval, pp. 541–548. Springer (2007)
11. Doshi-Velez, F., Wallace, B., Adams, R.: Graph-sparse IDA: a topic model with structured sparsity. arXiv:1410.4510 (2014)
12. Flisar, J., Podgorelec, V.: Document enrichment using dbpedia ontology for short text classification. In: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, pp. 1–9 (2018)
13. Garla, V.N., Brandt, C.: Ontology-guided feature engineering for clinical text classification. J. Biomed. Inform. **45**(5), 992–998 (2012)
14. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing? Int. J. Hum.-Comput. Stud. **43**(5–6), 907–928 (1995)
15. Hitzler, P., Krotzsch, M., Rudolph, S.: Foundations of Semantic Web Technologies. Chapman and Hall/CRC (2009)
16. Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: automatically evaluating topic coherence and topic model quality. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 530–539 (2014)
17. Manola, F., Miller, E., McBride, B., et al.: RDF primer. W3C Recomm. **10**(1–107), 6 (2004)
18. Mendes, P.N., Jakob, M., Garcia-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems (I-Semantics) (2011)
19. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108 (2010)
20. O'callaghan, D., Greene, D., Carthy, J., Cunningham, P.: An analysis of the coherence of descriptors in topic modeling. Exp. Syst. Appl. **42**(13),5645–5657 (2015)
21. Prud'hommeaux, E., Seaborne, A.: SPARQL query language for RDF. W3C Recommendation, W3C. Retrieved on 16 Nov 2009 (2008)
22. Rebele, T., Suchanek, F.M., Hoffart, J., Biega, J., Kuzey, E., Weikum, G.: YAGO: a multilingual knowledge base from wikipedia, wordnet, and geonames. In: The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, 17-2 Oct 2016, Proceedings, Part II, pp. 177–185 (2016). https://doi.org/10.1007/978-3-319-46547-019
23. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 399–408 (2015)
24. Steyvers, M., Griffiths, T.: Probabilistic topic models. Handb. Latent Seman. **427**(7), 424–440 (2007)

25. Suganya, G., Porkodi, R.: Ontology based information extraction-a review. In: 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), pp. 1–7. IEEE (2018)
26. Syed, S., Spruit, M.: Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. In: 2017 IEEE International conference on data science and advanced analytics (DSAA), pp. 165–174. IEEE (2017)
27. Vallet, D., Fernández, M., Castells, P.: An ontology-based information retrieval model. In: European Semantic Web Conference, pp. 455–470. Springer (2005)
28. Waitelonis, J.: Linked Data Supported Information Retrieval. Ph.D. thesis, Karlsruher Institut für Technologie (2018)
29. Yao, L., et al.: Incorporating knowledge graph embeddings into topic modeling. In: Thirty-first AAAI Conference on Artificial Intelligence (2017)