LOV-ES: Guiding the Ontology Selection to Structure Textual Data using Topic Modeling

Damien Graux, Anaïs Ollagnier

Inria, Université Côte d'Azur, CNRS, I3S, France

Abstract

On-line availability of text corpora nowadays allow data practitioners to build complex knowledge combining various sources. One common shared challenge lays in the modelisation of intermediate knowledge structures able to gather at once the various topics present in the texts. Practically, practitioners often go through the creation of vocabularies. In order to help these domain experts, we design LOV-ES: a solution able to help them in this creative process, guiding them in the selection and the combination of already existing vocabularies available online. Technically, our solution relies on LDA to detect topics and on the LOV to then propose candidate vocabularies.

Keywords

Topic Modeling, Textual source, LDA, LOV, Vocabulary Recommender

1. Introduction

Since its early days, natural language processing has *knowledge acquisition* as a prominent goal. Formally, we would like to have machines able to read text and express its embedded knowledge in a formal representation so to rely on it to later solve various problems. One of the first tasks to tackle in this context is therefore to have an ontology able to capture the knowledge structure of the considered texts: the *ontology induction*.

Ontology induction (constructing an ontology) and ontology population (mapping textual expressions to concepts and relations in the ontology) have been explored by the community [1]. In particular, early efforts have been made to bridge Semantic Web and ontology learning [2, 3]; and machine learning approaches were developed [4, 5], some being unsupervised [6]. Nevertheless, as highlighted by Tsujii [7], most approaches induce and populate a deterministic ontology, which does not capture the uncertainty among entities and relations. Moreover, they focus on inducing ontology over individual words rather than arbitrarily large meaning units.

In parallel, the Semantic Web community has been structuring many knowledge domains in an open manner, *i.e.* sharing their datasets and their ontologies, see *e.g.* the LOD-cloud[1] and the LOV-dataset[2] providing data on respectively open linked datasets and vocabularies. As a consequence, in this article, we propose to rely on these available resources to design

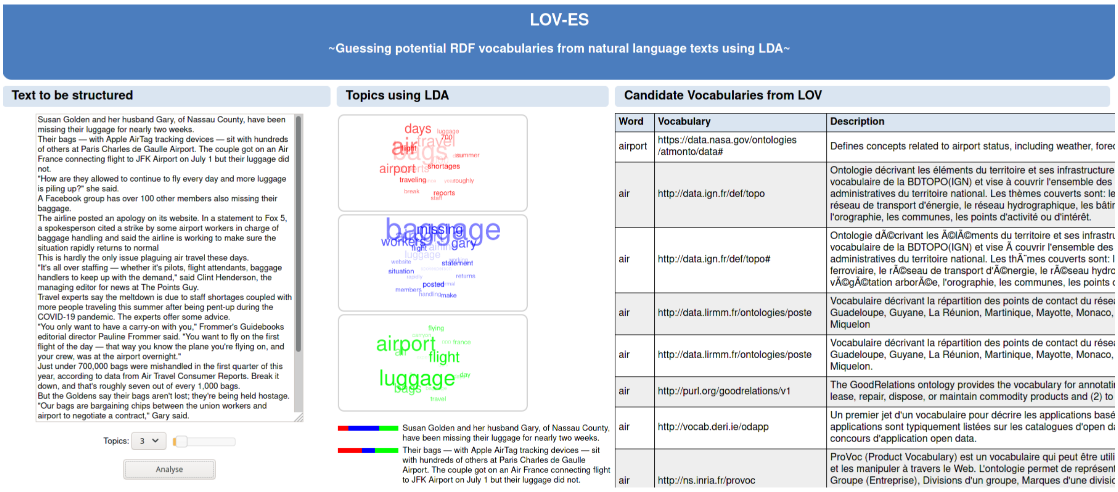
ISWC 2022: The 21st International Semantic Web Conference, 23-27 October 2022, Virtual (instead of Hangzhou, China)

damien.graux@inria.fr (D. Graux); anais.ollagnier@inria.fr (A. Ollagnier)
https://dgraux.github.io/ (D. Graux); https://aollagnier.github.io/ (A. Ollagnier)
0000-0003-3392-3162 (D. Graux); 0000-0003-4117-1191 (A. Ollagnier)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] LOD: more than 1 255 datasets. https://lod-cloud.net/

[2] LOV: more than 800 vocabularies. https://lov.linkeddata.es/dataset/lov



LOV-ES

~Guessing potential RDF vocabularies from natural language texts using LDA~

Text to be structured

Susan Golden and her husband Gary, of Nassau County, have been missing their luggage for nearly two weeks.
Their bags — with Apple AirTag tracking devices — sit with hundreds of others at Paris Charles de Gaulle Airport. The couple got on an Air France connecting flight to JFK Airport on July 1 but their luggage did not.
"How are they allowed to continue to fly every day and more luggage is piling up?" she said.
A Facebook group has over 100 other members also missing their baggage.
The airline posted an apology on its website. In a statement to Fox 5, a spokesperson cited a strike by some airport workers in charge of baggage handling and said the airline is working to make sure the situation rapidly returns to normal
This is hardly the only issue plaguing air travel these days.
"It's all over staffing — whether it's pilots, flight attendants, baggage handlers to keep up with the demand," said Clint Henderson, the managing editor for news at The Points Guy.
Travel experts say the meltdown is due to staff shortages coupled with more people traveling this summer after being pent-up during the COVID-19 pandemic. The experts offer some advice.
"You only want to have a carry-on with you," Frommer's Guidebooks editorial director Pauline Frommer said. "You want to fly on the first flight of the day — that way you know the plane you're flying on, and your crew, was at the airport overnight."
Just under 700,000 bags were mishandled in the first quarter of this year, according to data from Air Travel Consumer Reports. Break it down, and that's roughly seven out of every 1,000 bags.
But the Goldens say their bags aren't lost; they're being held hostage.
"Our bags are bargaining chips between the union workers and airport to negotiate a contract," Gary said.

Topics: 3

Topics using LDA

Susan Golden and her husband Gary, of Nassau County, have been missing their luggage for nearly two weeks.

Their bags — with Apple AirTag tracking devices — sit with hundreds of others at Paris Charles de Gaulle Airport. The couple got on an Air France connecting flight to JFK Airport on July 1 but their luggage did not.

Candidate Vocabularies from LOV

Word	Vocabulary	Description
airport	https://data.nasa.gov/ontologies /atmonto/data#	Defines concepts related to airport status, including weather, forec
air	http://data.ign.fr/def/topo	Ontologie décrivant les éléments du territoire et ses infrastructure vocabulaire de la BDTOPO(IGN) et vise à couvrir l'ensemble des administratives du territoire national. Les thèmes couverts sont: le réseau de transport d'énergie, le réseau hydrographique, les bâtir l'orographie, les communes, les points d'activité ou d'intérêt.
air	http://data.ign.fr/def/topo#	Ontologie dÃ©crivant les Ã©lÃ©ments du territoire et ses infrastru vocabulaire de la BDTOPO(IGN) et vise Ã couvrir l'ensemble des administratives du territoire national. Les thÃ¨mes couverts sont: l ferroviaire, le rÃ©seau de transport d'Ã©nergie, le rÃ©seau hydro vÃ©gÃ©tation arborÃ©e, l'orographie, les communes, les points c
air	http://data.lirmm.fr/ontologies/poste	Vocabulaire décrivant la répartition des points de contact du résea Guadeloupe, Guyane, La Réunion, Martinique, Mayotte, Monaco, Miquelon
air	http://data.lirmm.fr/ontologies/poste	Vocabulaire décrivant la répartition des points de contact du résea Guadeloupe, Guyane, La Réunion, Martinique, Mayotte, Monaco, Miquelon.
air	http://purl.org/goodrelations/v1	The GoodRelations ontology provides the vocabulary for annotati lease, repair, dispose, or maintain commodity products and (2) to
air	http://vocab.deri.ie/odapp	Un premier jet d'un vocabulaire pour décrire les applications basé applications sont typiquement listées sur les catalogues d'open da concours d'application open data.
air	http://ns.inria.fr/provoc	ProVoc (Product Vocabulary) est un vocabulaire qui peut être utili et les manipuler à travers le Web. L'ontologie permet de représent Groupe (Entreprise), Divisions d'un groupe, Marques d'une divisic

Figure 1: Screenshot of the LOV-ES interface.

a vocabulary recommender from text, aiming to guide and help data practitioners in their ontology selection and design process. Following a similar goal, domain-specific solutions have been designed *e.g.* for medical data [8].

In this article, we propose an end-to-end solution named LOV-ES for Linked Open Vocabulary Enhanced Selection. We show how LDA and SPARQL [9] can be associated to obtain a list of candidate vocabularies. Moreover, we propose a metric to rank the vocabularies against each other (combining words' weights and vocabulary frequencies in the list) in order to suggest the most relevant results.

2. Vocabulary Suggestion using LDA

The LOV initiative created in 2014 [10] has contributed to the deployment of Linked Data providing a high-quality catalogue of reusable vocabularies for the description of data on the Web. Currently, the LOV gathers around 800 vocabularies in more than 50 languages ranging from English to French. The task of Topic Modeling consists of discovering abstract semantic themes, or *topics*, hidden within data. Among existing topic model techniques, Latent Dirichlet Allocation (LDA) [11] and its extensions have been successfully applied to many data types and application domains, including bioinformatics, computer vision, and social network analysis, in addition to text mining and analytics [12].

The proposed architecture performs two main tasks: (1) applying LDA to identify underlying topics of a given text block, and (2) ranking candidate vocabularies considering their relevance to the topics. In detail, we rely on the original version of LDA [11] based on two hyperparameters α (controlling the prior distribution over topic weights in each document) and β (setting the prior distribution over word weights in each topic), respectively set to 0.1 and 0.01. Each topic word bag resulting from applying LDA is then used to build a SPARQL query aiming at extracting candidate vocabularies. Typically, the words are given to the SPARQL endpoint taking advantage of the VALUES variable passing method as presented schematically below:

```
SELECT ?word ?voc WHERE {
  VALUES ?word { "word1" "word2" "word3" } # Word list from an LDA bag.
  ?voc a lod:Vocabulary . ?voc dcterms:description ?description . [...]
  FILTER ( CONTAINS ( STR(?description),?word ) )   }
```

The rest of the query then returns candidate vocabularies ?voc if occurrences of ?word are present within the vocabulary description, making the assumption that specific words would be present to describe more generic vocabularies.

Once candidate vocabularies are obtained. We apply the following metric to give each vocabulary a score of relevancy. Formally, considering k topics, we have for each topic t a bag containing n_t words and their associated weights *i.e.* $\{(w_{i,t}, p_{i,t}\}_t$, with $i \in [1, n_t]$ and $t \in [1, k]$. As there might be multiple possible candidate vocabularies for a single word, we obtained (after running the k SPARQL queries) k lists containing c_t 3-tuples of the form $(voc_j, w_{i,t}, p_{i,t})$, with $j \in [1, c_t], i \in [1, n_t]$ and $t \in [1, k]$. Moreover, a specific vocabulary can be suggested multiple times for a single topic *e.g.* a vocabulary about travel may be returned two times considering the bag {“plane”,“trip”}. To rank the list, we propose to first attribute an aggregated weight for each vocabulary per topic, summing the weights of the words it matches while normalising by the sum of the weights in the 3-tuple list for the considered topic; and then to combine the topics summing the aggregated weights of same vocabularies together:

$$\forall V \in vocList, \quad score(V) = \sum_{t=1}^{k} \frac{\sum\limits_{voc_i=V} p_{i,t}}{\sum\limits_{i=1}^{c_t} p_{i,t}}, \quad \text{considering } k \text{ lists of } c_t \text{ 3-tuples}$$

We then just have to decreasingly sort the $score$ of each vocabulary of $vocList$ to provide an ordered list of suggestions.

The graphical user interface takes as input a text block split by line and allows to select the number of topics to be identified, ranging from 2 to 10. By clicking on the *Analyse* button, word clouds are generated corresponding to detected topics, graphically the larger a word, the more weight it has in the topic. In parallel, a ranking list of candidate vocabularies is displayed by descending order of relevancy (using the scoring function aforementioned) including among other information: the link to the vocabulary, its description and various explanations to describe the internal step of the process. In particular, it provides information referring to the distribution of words over topics after applying the scoring metric. Figure 1 presents our pure JavaScript Web-App: LOV-ES.

3. Related Work

Most of existing works focusing on the task of ontology research relies on the content of ontology and user query to perform vocabulary recommendation [13]. In these works, the core idea consists of using a set of keywords or ontology metadata to describe the domain to suggest appropriate ontology according to the user query. Swoogle [3], Sindice.com [4] or Watson [5] are semantic search engines performing a search of ontology resources using the aforementioned methodology. Among proposed works, some of them have especially focused on the LOV ecosystem. Initially consisting of a full-text inverted index and a ranking algorithm based on the term popularity [10], related works have extended the recommendation capabilities of the LOV search engine proposing ranking metrics [14, 15, 16] and evaluation parameters [17]. Recently, Sarwar et al. [18] have addressed the task as an information retrieval problem, introducing a framework-based on text categorisation and unsupervised learning techniques. To overcome the well-known limitations related to domain coverage and keyword-based searching/matching, LOV-ES benefits from NLP-powered techniques allowing it to consider full text data to performing ontology resource recommendation.

4. Conclusion

This article presents LOV-ES: a solution combining an LDA-based model together with a novel scoring metric to extract relevant vocabularies from the LOV catalog in order to structure a text. A pure JavaScript Web-App is openly available from:

https://dgraux.github.io/LOV-ES/

with the aim of guiding data practitioners selecting relevant vocabularies to structure textual knowledge. Currently working exclusively with resources released on LOV, it constitutes the main limitation of the proposed solution. In future work, we plan to extend the coverage of available vocabularies by aggregating resources from other catalogs. In addition, several approaches for estimating LDA parameters have been proposed in the literature and should be considered in an improved version of LOV-ES, together with an evaluation of the results.

Acknowledgments. This work is funded under the IDEX UCA OTESIA, the UCA Academy 1 project (C870A021–D103–ACAD1_FIN_17_20Y) and the French government, through the 3IA Côte d'Azur Investments managed by the National Research Agency (ANR-19-P3IA-0002).

References

[1] S. Staab, R. Studer, Handbook on ontologies, Springer Science & Business Media, 2010.

[2] P. Cimiano, Ontology learning and population from text: algorithms, evaluation and applications, volume 27, Springer Science & Business Media, 2006.

[3] A. Maedche, S. Staab, Ontology learning for the semantic web, IEEE Intelligent systems 16 (2001) 72–79.

[4] R. Snow, D. Jurafsky, A. Y. Ng, Semantic taxonomy induction from heterogenous evidence, in: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, 2006, pp. 801–808.

[3] https://ebiquity.umbc.edu/project/html/id/53/Swoogle

[4] https://sindice.com/

[5] http://watson.kmi.open.ac.uk

[5] F. Wu, D. S. Weld, Automatically refining the wikipedia infobox ontology, in: Proceedings of the 17th international conference on World Wide Web, 2008, pp. 635–644.

[6] H. Poon, P. Domingos, Unsupervised ontology induction from text, in: Proceedings of the 48th annual meeting of the Association for Computational Linguistics, 2010, pp. 296–305.

[7] J. Tsujii, Thesaurus or logical ontology, which do we need for mining text?, in: LREC, 2004.

[8] M. Martínez-Romero, C. Jonquet, M. J. O'connor, J. Graybeal, A. Pazos, M. A. Musen, Ncbo ontology recommender 2.0: an enhanced approach for biomedical ontology recommendation, Journal of biomedical semantics 8 (2017) 1–22.

[9] S. Harris, A. Seaborne, E. Prud'hommeaux, Sparql 1.1 query language, W3C recommendation 21 (2013) 778.

[10] P. Vandenbussche, G. Atemezing, M. Poveda-Villalón, B. Vatant, Linked open vocabularies (LOV): A gateway to reusable semantic vocabularies on the web, Semantic Web 8 (2017) 437–452. URL: https://doi.org/10.3233/SW-160213. doi:10.3233/SW-160213.

[11] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

[12] I. Vayansky, S. A. P. Kumar, A review of topic modeling methods, Inf. Syst. 94 (2020) 101582. URL: https://doi.org/10.1016/j.is.2020.101582. doi:10.1016/j.is.2020.101582.

[13] A. S. Butt, A. Haller, L. Xie, DWRank: Learning concept ranking for ontology search, Semantic Web 7 (2016) 447–461.

[14] G. A. Atemezing, R. Troncy, Information content based ranking metric for linked open vocabularies, in: H. Sack, A. Filipowska, J. Lehmann, S. Hellmann (Eds.), Proceedings of the 10th International Conference on Semantic Systems, SEMANTiCS 2014, Leipzig, Germany, September 4-5, 2014, ACM, 2014, pp. 53–56. URL: https://doi.org/10.1145/2660517.2660533. doi:10.1145/2660517.2660533.

[15] I. Stavrakantonakis, A. Fensel, D. Fensel, Linked open vocabulary ranking and terms discovery, in: A. Fensel, A. Zaveri, S. Hellmann, T. Pellegrini (Eds.), Proceedings of the 12th International Conference on Semantic Systems, SEMANTiCS 2016, Leipzig, Germany, September 12-15, 2016, ACM, 2016, pp. 1–8. URL: https://doi.org/10.1145/2993318.2993338. doi:10.1145/2993318.2993338.

[16] N. Kolbe, P. Vandenbussche, S. Kubler, Y. L. Traon, Lovbench: Ontology ranking benchmark, in: Y. Huang, I. King, T. Liu, M. van Steen (Eds.), WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020, ACM / IW3C2, 2020, pp. 1750–1760. URL: https://doi.org/10.1145/3366423.3380245. doi:10.1145/3366423.3380245.

[17] N. Kolbe, S. Kubler, J. Robert, Y. L. Traon, A. B. Zaslavsky, Linked vocabulary recommendation tools for internet of things: A survey, ACM Comput. Surv. 51 (2019) 127:1–127:31. URL: https://doi.org/10.1145/3284316. doi:10.1145/3284316.

[18] M. A. Sarwar, M. Ahmed, A. Habib, M. Khalid, M. A. Ali, M. Raza, S. Hussain, G. Ahmed, Exploiting ontology recommendation using text categorization approach, IEEE Access 9 (2021) 27304–27322. URL: https://doi.org/10.1109/ACCESS.2020.3047364. doi:10.1109/ACCESS.2020.3047364.