Learn2Construct: An automatic ontology construction based on LDA from texual data.

Ahmed Khemiri
VPNC Lab., FSJEGJ, University of Jendouba
Jendouba, Tunisia
ahmedkhemiri24@outlook.fr

Amani Drissi
VPNC Lab., FSJEGJ, University of Jendouba
Jendouba, Tunisia
drissiamani19@gmail.com

Anis Tissaoui
VPNC Lab., FSJEGJ, University of Jendouba
Jendouba, Tunisia
anis.tissaoui@fsjegj.rnu.tn

Salma Sassi
VPNC Lab., FSJEGJ, University of Jendouba
Jendouba, Tunisia
salma.sassi@fsjegj.rnu.tn

Richard Chbeir
Univ Pau & Pays Adour, E2S UPPA, LIUPPA
Anglet, France
richard.chbeir@univ-pau.fr

ABSTRACT

In recent years, the research on Ontology Learning has become a hot topic among researchers because of the exponential increase of the number of documents and textual data not only on the web but also in digital libraries. This has participated to the emergence of new computational tools and methods to deal with the automatic organization, representation, retrieval and exploration of large corpus in order to have a good way of organizing and managing huge volumes of data. LDA-based approaches have proven to provide the best result [18][16] [4]. However, they suffers to several limitations related to concept and relation extraction, as well as coping with the corpus evolution. In order to cope with these problems, we propose here a new solution named Learn2Construct which is an automatic ontology construction method based on topic modeling. Experiments have been conducted to measure the effectiveness of our solution and compare it to existing ones. The results obtained are more than satisfactory.

CCS CONCEPTS

• **Information systems → Document topic models**.

KEYWORDS

Ontology Learning, Topic modeling, Text classification, Latent Dirichlet Allocation, LDA, Ontology based on topic modeling.

ACM Reference Format:
Ahmed Khemiri, Amani Drissi, Anis Tissaoui, Salma Sassi, and Richard Chbeir. 2021. Learn2Construct: An automatic ontology construction based on LDA from texual data.. In *International Conference on Management of Digital EcoSystems (MEDES '21), November 1–3, 2021, Virtual Event, Tunisia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3444757.3485110

MEDES '21, November 1–3, 2021, Virtual Event, Tunisia
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8314-1/21/11…$15.00
https://doi.org/10.1145/3444757.3485110

1 INTRODUCTION

Ontologies are widely used for knowledge representation. They play a major role in supporting the information exchange and sharing. Their usage has proven to be beneficial and efficient in different applications (e.g., information retrieval, information extraction, and question answering). Ontology learning (OL) from text is a process aiming at automatically extracting and representing the knowledge from text in machine-readable form [2]. It refers to extracting conceptual knowledge from several sources and building ontology from scratch, or enriching / adapting an existing ontology. The vision of OL includes a number of complementary disciplines that feed on different types of unstructured, semi-structured and fully structured data in order to support an automatic ontology engineering process.

In recent years, the research on OL becomes a hot topic among researchers due to the exponential increase of the number of documents and textual data not only on the web but also in digital libraries. This has participated to the emergence of new computational methods and tools to deal with the automatic organization, representation, retrieval and exploration of a large corpus. To this end, several approaches have been elaborated [21], [5], [14], [8], [9]. Most of them used topic modeling algorithms such as LDA (Latent Dirichlet Allocation), LSA (Latent Semantic Analysis) and PLSA (Probabilistic Latent Semantic Analysis) to automate the text classification process. In -[13], [17], [15] and [12], the authors build a domain ontologies in order the facilitate the reuse of data and clarify the representation of a large corpus. In [22], [18], [7] and [11], the aim was to build a domain ontology based on topic modeling algorithms in order to guarantee a better document text classification and to better represent the document knowledge. It is to be mentioned that the Latent Dirichlet Allocation or LDA-based topic modeling algorithms are mostly used in the literature. This is due to LDA successful processing since: 1) it works well with large corpus, 2) it does not suffer from overfitting issues like in PLSA, and 3) it can be embedded in other complex models. However, the majority of existing LDA-based works (e.g., [18] and [7]) can build an ontology only after the acquisition of learning data produced by LDA including topics and their associated vocabularies (bottom-up principle) which generates inconsistencies due to corpus change in size or content. For the purpose of constructing, enriching and using efficiently ontologies, there are many challenges to be addressed beforehand:

(1) How to automatically extract relevant terms from a heterogeneous corpus and classify them by topic?
(2) How to identify from the extracted candidate terms (classified by topic), the concepts and the relationships between them?
(3) How to automatically re-stabilize the learning model in order to guarantee its performance even when the size of the corpus is evolving?

In order to overcome these challenges, we provide here a new approach consisting of automatic construction of an ontology from a textual corpus using LDA unsupervised algorithm. Our approach is able to **(i)** automatically extract and classify terms from heterogeneous corpus according to their dominant topics, **(ii)** identify candidate concepts and deduce semantic relationships between them, and finally **(iii)** automatically stabilize the construction process in order to ensure model stability regardless of the corpus changing.

The rest of the paper is organized as follows. We describe in Section 2 the ontology learning process as well as the topic modeling algorithms. We also describe and discuss the existing works of Ontology Construction from Texts based on topic modeling algorithms. Next, we illustrate in Section 3 our automatic ontology construction based on LDA (Learn2Construct) methodology. Section 4 describes the experiments conducted to validate our approach. It also shows the comparison of our results with an existing method based on LDA on the same dataset. Section 5 focuses on the evaluation of results. Finally, Section 6 concludes the paper and discusses some future work.

2 LITERATURE REVIEW

2.1 Background

Ontology learning is the most important move towards reducing the cost of ontology building. It refers to the bundle of semi-automatic frameworks applied to ontology construction and maintaining [1]. OL concerns knowledge acquisition which is a multidisciplinary field using multiple techniques: semantic web, ML (Machine Learning), logic, knowledge representation, philosophy, databases, NLP (Natural language processing), IR (information retrieval), reasoning and AI (Artificial intelligence) [10]. Over the last few years, some efforts have been made to build semi-automatic ontology construction tools [19]. Topic modeling[1] is the new revolution in text mining and successfully used in ontology learning. It is a statistical technique used to analyse a huge volume of data and to extract hidden concepts, prominent feature or latent variables of data, depending on the application context citePooja+2018. Topic modeling generate topics. Each topic contains a cluster of words that frequently occur together [3]. The most popular topic modeling algorithms used for text analysis in several domains includes Latent Semantic Analysis[2], Probabilistic Latent Semantic Analysis[3] and Latent Dirichlet Allocation[4]. It is to be mentioned that the Latent Direchlet Allocation or LDA-based topic modeling algorithms are mostly used in the literature. This is due to LDA successful processing since: 1) it works well with large corpus, 2) it does not suffer from overfitting issues like in PLSA, and 3) it can be embedded in other complex models.

2.2 Related work

The automatic construction of ontology from text consists in analyzing the collected text, identifying terms, constructing concepts hierarchies, identifying relationships between concepts and evaluating the constructed ontology.

In this context, several works relating to the automatic or semi-automatic domains ontology construction from texts like [13] are proposed in the literature. these approaches are based on machine learning algorithms.

[17] constructs an automatic ontology in order to enrich the existed domain ontology using unstructured data. The main objective of this work is to extract semantic relations based on a supervised model that induces symbolic rules for extracting binary relations between entities from textual corpora. [15] creates an automatic domain ontology from unstructured data in order to identify the domain concepts as well as their hierarchical and non-hierarchical relationships. [15] applied the HITS algorithm in the objective to successfully extract the most important domain concepts. For the hierarchical relations extraction (Hypernym-Hyponym relations) the author used the Morpho Syntactic Pattern. The non-hierarchical relations were extracted by applying a rule based method.

In [12], the author builds an automatic domain ontology from unstructured data, by extracting the domain concepts using C/NC-value method as well as their hierarchical and non-hierarchical relationships. The hierarchical relationships were extracted using Hierarchical clustering and Formal Concept Analysis, the experiments discussed towards the end of this work showed that Hierarchical clustering is more efficient than Formal Concept Analysis in this task. For the non-hierarchical relations extraction, [12] applied two methods which are the Association Rules and a Probabilistic algorithm.

Several researches based on Topic modeling techniques have been also proposed to automate the ontology construction, such as ([21]; [5], [14]; [8] and [9] using topic modeling to classify their documents according to topics using Latent Dirichlet Allocation algorithm (LDA), Latent Semantic Analysis algorith (LSA) or Probabilistic Latent Semantic Analysis algorithm (PLSA).

To analyze existing works and with the absence of common measures, we define here six criteria used to compare and discuss existing studies based on LDA unsupervised algorithm.

Challenge1: How to automatically extract relevant terms from a heterogeneous corpus and classify them by topic?

- **Criterion 1 (C1):** The Type of knowledge resources for which to learn an ontology: (i) Structured data such as existing ontologies and database schema. (ii) Semi-structured data like XML or HTML documents, web pages. (iii) Unstructured data including textual data.
- **Criterion 2 (C2):** The Term Type: (it can be 'simple' consisting of just one word, or 'complex' consisting of two or more words).

[1] https://monkeylearn.com/blog/introduction-to-topic-modeling/
[2] http://lsa.colorado.edu/papers/dp1.LSAintro.pdf
[3] https://arxiv.org/ftp/arxiv/papers/1301/1301.6705.pdf
[4] http://alberto.bietti.me/files/rapport-lda.pdf

	Challenge 1		**Challenge 2**		**Challenge 3**			
Criterion/Existing study	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**
Anoop[6]	Semi-structured	Complex	Term Concept	Concepts hierarchy	Automatic	Specific	No	User evaluation
Zhu [22]	Unstructured	Complex	Term	-	Semi-Automatic	Specific	Perplexity	User evaluation
Monika [18]	Unstructured	Simple	Term	-	Semi-Automatic	Specific	No	No
Attigeri [7]	Unstructured	Simple	Term	-	Semi-Automatic	Specific	No	No
Chowdhry [11]	Unstructured	Complex	Term	-	Semi-Automatic	Specific	No	No
Learn2Construct	Unstructured	Complex	Term Concept Non-hierarchical relations	Non-hierarchical relations	Automatic	Generic	Coherence	User evaluation Task-Based-evaluation

Table 1: Comparison of existing approaches and the proposed one

Challenge 2: How to identify from candidate terms (classified by topic), concepts and relationships between them?

- **Criterion 3 (C3):** Considered Tasks, the ontology learning task that be considered in the ontology construction process ('Terms', 'Concept', 'Relations' or 'Axioms').
- **Criterion 4 (C4):** Relations Types: (including 'Hierarchical Relations' and 'Nun-hierarchical Relations').

Challenge 3: How to automatically re-stabilize our model in order to guarantee its performance even with the change of the domain or the corpus size?

- **Criterion 5 (C5):** The level of automation: i) Semi-automatic if there is a user intervention in the ontology construction process, and ii) Automatic with no human intervention.
- **Criterion 6 (C6):** The Application domain: it can be 'Generic/ domain independent' or 'Specific/ particular domain.
- **Criterion 7 (C7):** The Model evaluation: the model evaluation can be established with the use of 'Perplexity' or 'Coherence' measures.
- **Criterion 8 (C8):** Ontology evaluation: it can be carried out with the use of these criteria: i) Task-Based Evaluation: an evaluation measure choice depends on the particular task. ii) Corpus-Based Evaluation: it is used for domain-specific data sources to check the coverage scope level of ontologies for a corresponding domain. iii) Criteria-Based Evaluation: it evaluates ontology by characterizing how nicely these follow a set of criteria. iv) User Evaluation: it is carried out by one or more human experts who have to assess to what extent the information obtained is correct. And v) Gold Standard Evaluation: it compares the learned ontology with predefined gold standard ontology. The gold standard ontology is generally manually created from scratch by domain experts.

Our analysis demonstrates that all existing works used unstructured data to build their ontologies except [6] who used semi-structured data (HTML documents) as input to build domain ontology. Also, table 3 illustrates that all existing studies used a semi-automatic domains-specific approach.

We observed also that in the extraction phase, the most of existing systems rely only on terms and specially on simple ones ([18] [7]). Only ([22] [11] and [6] extract complex terms but they do not evaluate the relevance of the extracted ones.

In the classification phase, we also note that most of existing studies focus only on terms and do not consider the other elements (concepts, relations, axioms) expect [6] who extracted the domain concepts using linguistic patterns. However, authors do not evaluate the relevance of their extracted concepts which lead consequently to produce syntactic and semantic inconsistency.

In the concept hierarchy construction phase, few works identify relations between concepts and we note that existing works extract either taxonomic relationships or non-taxonomic relationships conducting to generate semantic inconsistency. For example, [6] extracted the concepts hierarchy only using rule-based method, the other existing studies did not go beyond the hierarchy construction.

According to the evaluation phase, we also note that existing works did not evaluate the performance of the topic modeling except [22] who used the perplexity measure to evaluate the performance of the training model. However, the perplexity measure still enable to ensure model stability even changing the size of the corpus.

Finally, we note that existing works do not evaluate the ontology constructed according to the different evaluation methods detailed in the literature except ([22] and [6] who evaluated their ontologies with a user based evaluation.

To conclude our discussion, we note the absence of an approach taking into account in the same framework all the ontology construction process including the terms extraction, the concept identification, the hierarchy construction and the evaluation of the constructed ontology.

The main contribution of the paper consists of developing a new automatic method for building ontology from the extraction phase to the evaluation phase based on the probabilistic topic model of the

LDA. One strong aspect of our contribution is the use of topic models to build ontology. Our method named Learn2Construct expands the applicability of topic models to automated ontology construction. Experimental and evaluation have shown the soundness and the appropriateness of Learn2Construct method.

3 METHODOLOGY

Our ontology LDA-based construction learning method, called Learn2Construct, uses text information as data source to automatically create ontology concepts and inter-relationships. Learn2Construct framework is shown in Figure 1. The proposed architecture consists of five modules: **(A)** Preprocessing, **(B)** Terms extraction, **(C)** Topic modeling, **(D)** Concepts & Relations extraction, and **(E)** Ontology visualization modules. We detail them in what follows.

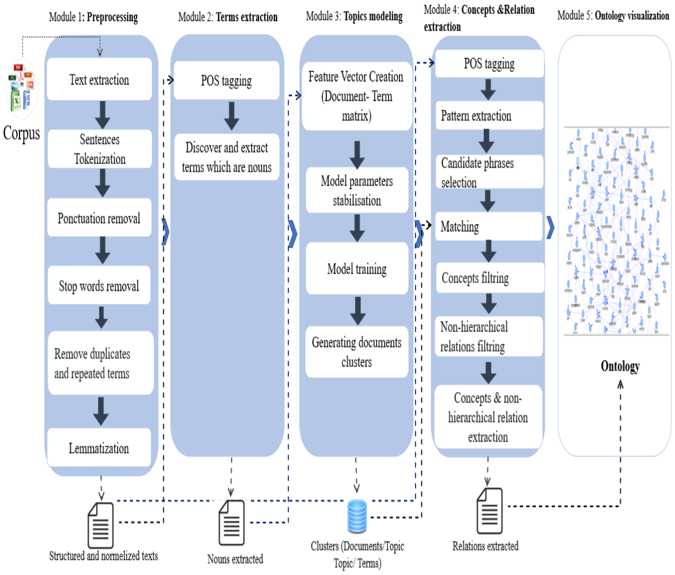

Module 1: Preprocessing
Module 2: Terms extraction
Module 3: Topics modeling
Module 4: Concepts &Relation extraction
Module 5: Ontology visualization
Corpus
Text extraction
Sentences Tokenization
Ponctuation removal
Stop words removal
Remove duplicates and repeated terms
Lemmatization
POS tagging
Discover and extract terms which are nouns
Feature Vector Creation (Document- Term matrix)
Model parameters stabilisation
Model training
Generating documents clusters
POS tagging
Pattern extraction
Candidate phrases selection
Matching
Concepts filtring
Non-hierarchical relations filtring
Concepts & non-hierarchical relation extraction
Ontology
Structured and normelized texts
Nouns extracted
Clusters (Documents/Topic Topic/ Terms)
Relations extracted

Figure 1: Learn2Construct framework.

3.1 Preprocessing

The preprocessing module consists of cleaning and normalizing input texts. It consists of: **(i)** Sentence tokenization is used to segment a text into sentences in order to facilitate the linguistic study. Tokens are very useful for finding such patterns and considered as a base step for stemming and lemmatization; **(ii)** Part of speech is used to tag the verbs and the nouns in the sentences; **(iii)** NP Chunking is used to extract the compound terms; and **(iv)** Normalization consists of two steps:

- Transform a text into lowercase, and
- Remove all details between square, brackets and parentheses (e.g., non-ASCII, punctuation, lemmatization, stop words, duplicates and repeated terms).

3.2 Term Extraction

In this module, adaptive word segmentation is carried out to extract relevant words. In our study, we used an NLP-based method (e.g., NLTK or Natural Language Processing Toolkit) but other methods can be used as well. At first, we identify the domain vocabulary automatically, and add it to the word segmentation dictionary. Then, we traverse the collection of each document in the word segmentation operation, to achieve adaptive word segmentation in order to form the domain terms set. To improve the training efficiency of LDA subject model and enhance the quality of ontology learning, we need to filter out unnecessary words to form candidate term set of OL. It consists of stop word filtering, low frequency word filtering and part-of-speech filtering in terms of domain term set. After the above process, a synonym words list is generated. This list can be reviewed by experts and considered as “correctly generated” with no errors. So, a candidate term set of a domain ontology learning is formed, which provides the training sample data for LDA subject model.

3.3 Topic Modeling

In this module, an LDA model is built. The model is trained by candidate term sets so to infer the semantic related terms automatically. The LDA model assumes that documents are mixtures of topics, while topics are probability distributions over the vocabulary. When the topic proportions of documents are judged, they can be exploited as the themes (high level representations of the semantics) of the documents. Topic models, which frequently represent topics as multi-nominal distributions over words, have been widely used for discovering latent topics in text collection. The proposed model consists of automatically generating terms related to each topic, to determine the dominant topic of each document and the dominant terms for each topic using LDA. Our module is based on:

(1) **Feature Vector Creation:** we convert all texts to a digital representation which are numeric vectors, since learning algorithms and classifiers cannot directly process textual data in its original form.
(2) **Model Parameters Stabilization:** the objective of this step is to stabilize the LDA model. The stabilization process is based on alpha (the hyper parameter indicating the document-topic distribution) and beta (the hyper parameter indicating the topic-term distribution) values as well as the topics number. So, to have a stable and efficient model, it is necessary to choose the optimal combination of alpha and beta values by taking into consideration the optimal number of topics according to a corpus. To achieve this, we used the coherence parameter (the Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic) in order to correctly stabilize our model.
(3) **Model Training:** we implemented our approach using the unsupervised algorithm Latent Dirichlet Allocation (LDA) topic modeling by optimizing the algorithm with the best combination of the optimal values of alpha and beta as well as the optimal number of topics based on the coherence measure.
(4) **Generating documents clusters:** this step consists in generating and viewing the documents/topics clusters.

Figure 2 shows the link between the corpus documents and their topics, as well as the terms representing each topic.

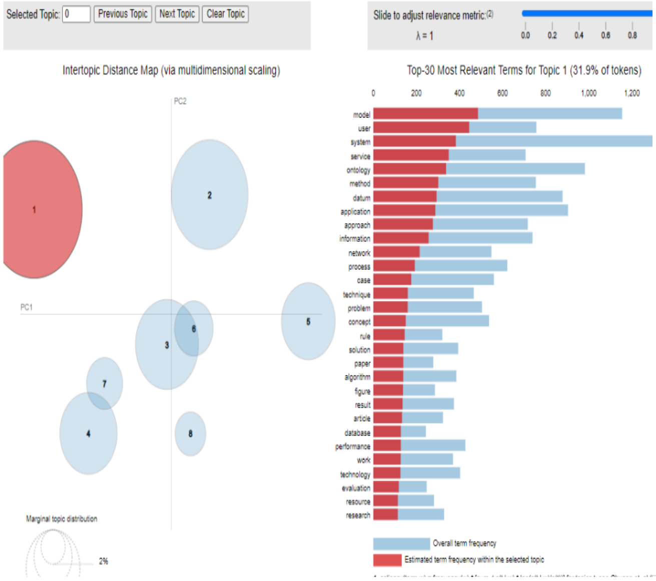


Figure 2: Documents topics generation.

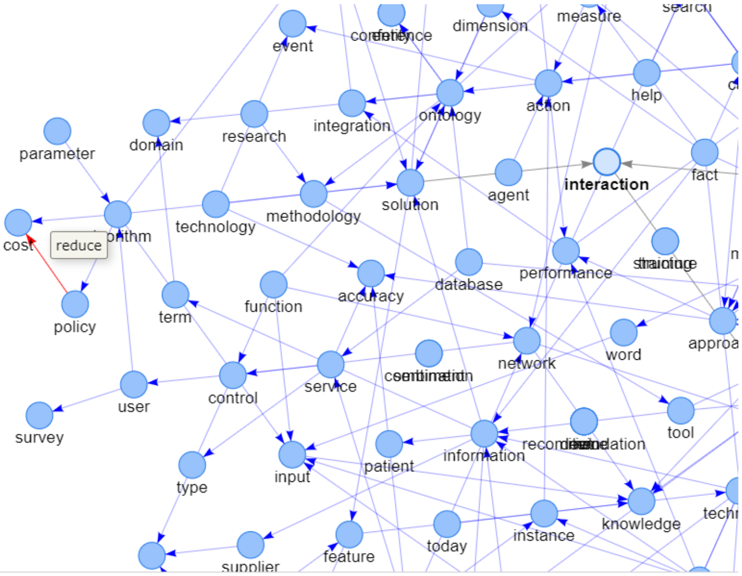


Figure 3: Example of concepts extracted with their relationships.

3.4 Concepts & Relation Extraction

Identifying relationships is very useful for automatic ontology bu ing. In our approach, we extract semantic relations using patte (Noun, Verb, Noun). This technique is applied on the internal l of the sentence to determine the semantic relations between concepts of our ontology with success. The first step consist applying the Part-Of-Speech. Then, we apply the proposed pat (noun, verb, noun) on the preprocessed corpus. Afterwards, extract all the candidate sentences by checking for each sente whether the couple of concepts belong to the same topic or Remember that the terms of each topic generated by LDA are related which guarantees their semantic similarity. The next ste to filter the candidate sentences and to match them with the se the candidate terms generated by LDA in order to determine concepts (in our case, the concept is each term which has a rela with another term, by checking their belonging to the same to of our ontology as well as their semantic relationships.

Figure 3 illustrates the relationship 'reduce' between the two cepts policy and cost.

3.5 Ontology Visualization

In order to better represent and to graphically visualize our or ogy, we implemented an interactive visualization (Figure 4), wl tries to answer the basic notions:

(1) What are the concepts of our ontology?
(2) How are the concepts related to each other?
(3) What are the relationships that can link them together?

This interactive interface contains the concepts of the corpus as well as their semantic relationships.

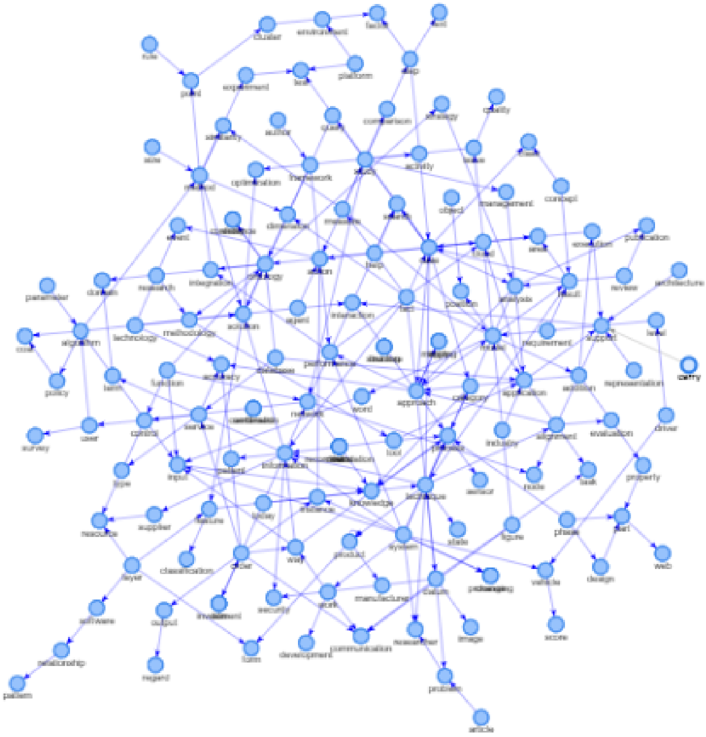


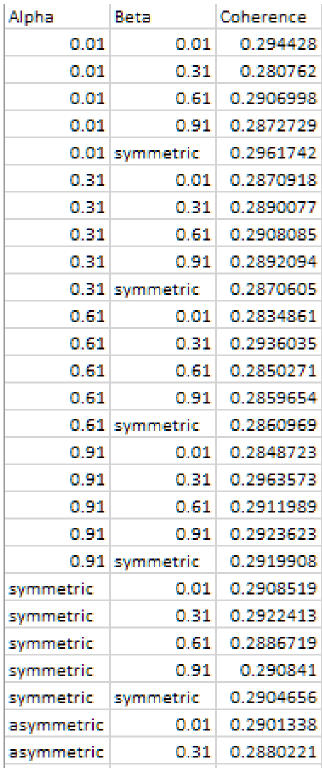
Figure 4: Implemented Ontology Graph Visualization.

4 EXPERIMENTS

Our experiments aimed to evaluate: **(i)** the quality of the constructed ontology **(objective 1)** and **(ii)** the performance of our proposed method comparing to existing ones and its applicability to other domains **(Objective 2)**. In the following, we detail the experimental protocols used to achieve these objectives. After, we detail the evaluation process and results.

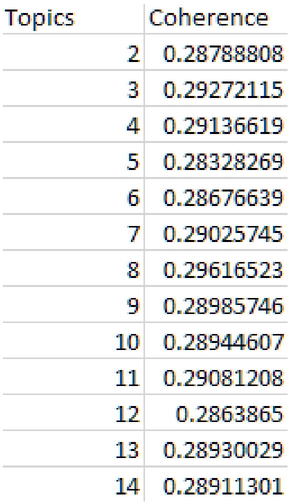
4.1 Experimental Protocols

4.1.1 Experimental Protocol 1. We have generated a set of scientific papers (corpus 1 and corpus 2) containing more than 904 thousand words (corpus 1). To guarantee a higher performance of our model, we have worked on the automation of our LDA model stabilization within the objective to choose the best alpha and beta values based on the coherence measure (defined in the next subsection). Thus, we have automatically tested the model with all the possible combinations of alpha and beta default values to ensure that we have chosen the best one. Figure 5 shows that the best combination with our corpus is 0.91 for alpha value and 0.31 for beta value. These values are automatically chosen by our model because they are relative to the highest coherence value. The performance of the LDA model also depends on the number of topics. Figure 6 shows that the results obtained when varying the number of topics. Once can see that the best result is obtained when the number of topics is 8. In order to evaluate the performance of our approach, we have implemented it with a larger textual corpus, which is built using published papers which note 15201 thousand words (corpus 2) (see section 4) .



Alpha	Beta	Coherence
0.01	0.01	0.294428
0.01	0.31	0.280762
0.01	0.61	0.2906998
0.01	0.91	0.2872729
0.01	symmetric	0.2961742
0.31	0.01	0.2870918
0.31	0.31	0.2890077
0.31	0.61	0.2908085
0.31	0.91	0.2892094
0.31	symmetric	0.2870605
0.61	0.01	0.2834861
0.61	0.31	0.2936035
0.61	0.61	0.2850271
0.61	0.91	0.2859654
0.61	symmetric	0.2860969
0.91	0.01	0.2848723
0.91	0.31	0.2963573
0.91	0.61	0.2911989
0.91	0.91	0.2923623
0.91	symmetric	0.2919908
symmetric	0.01	0.2908519
symmetric	0.31	0.2922413
symmetric	0.61	0.2886719
symmetric	0.91	0.290841
symmetric	symmetric	0.2904656
asymmetric	0.01	0.2901338
asymmetric	0.31	0.2880221

Figure 5: Coherence values when varying the alpha and beta values using corpus 1.



Topics	Coherence
2	0.28788808
3	0.29272115
4	0.29136619
5	0.28328269
6	0.28676639
7	0.29025745
8	0.29616523
9	0.28985746
10	0.28944607
11	0.29081208
12	0.2863865
13	0.28930029
14	0.28911301

Figure 6: Coherence values when varying the topics number using corpus 1.

4.1.2 Experimental Protocol 2. In order to assess the performance of our approach compared to the existing studies, we used wiki documents to build our datasets:

- Corpus 3: composed of 200,000 wiki documents.
- Corpus 4: composed of 950,680 wiki documents.

This evaluation consists of using our approach "Learn2Construct" in order to build an automatic ontology based on these two corpora. After that, we used tool provided by [6] for the same objective, the choice of this method to evaluate our model is not randomly, most of the works based on topic modeling algorithms have built their ontologies with only terminology extraction, on the other hand [6] is the only one that extracts the concepts and relationships between them to build his ontology so it is the most similar method compared to our approach. Then, we evaluated the achieved results using DBpedia ontology. This evaluation is based on two criteria which are the number of relevant concepts as well as the number of correctly identified relationships validated by DBpedia based on the precision measure (defined in the next subsection).

4.2 EVALUATION

The first experiments are evaluated using:

- **Criteria based evaluation:** in order to evaluate the performance of our approach, we used the coherence measure which scores a single topic by measuring the degree of semantic similarity between high scoring words in the topic. We generalize this as

$$Coherence(V) = \sum_{(V_i,\, V_j) \in V} Score_{(V_i,\, V_j,\, e)} \tag{1}$$

where V is a set of word describing the topic and e indicates a smoothing factor which guarantees that score returns real numbers [20].

- **Task based Evaluation:** which is a quantitative evaluation using conventional measures in information retrieval such as recall, precision, and F-measure, denoted as R, P, and Fscore respectively. The Recall R is defined as:

$$R = \frac{C_{sc}}{K_{sc}} \tag{2}$$

where C_{sc} defines the number of correct learned statements and K_{sc} defines the number of correct statements.
The Precision P is defined as:

$$P = \frac{Co_{sc}}{Id_{sc}} \tag{3}$$

where Co_{sc} is the number of correct learned statements, Id_{sc} is the total number of learned statements.
To assess the performance of our ontology learning method, we note that precision measure alone is not sufficient. Indeed, low recall means that a large collection of topics is not selected. The F-score measure (or F1) is defined as the harmonic mean of recall and precision:

$$F - score = \frac{2 \times P \times R}{P + R} \tag{4}$$

Figure 7 shows that Learn2Construct has a precision of 81% , a recall of 79% and an F-score of 79.98 % using corpus 1. It also shows that Learn2Construct has a precision of 95 % , a recall of 90% and an F-score of 92.43% using a larger corpus which is constituted with the corpus 2. This evaluation shows that our approach is more efficient with a larger corpus (corpus2). Table 2 shows the results of evaluating Learn2Construct and [6] approach for the concepts extraction using the Corpus 3. With the application of Learn2Construct approach, we succeeded the extraction of 5145 concepts of which 4680 are relevant according to DBpedia. On the other hand, when we adapted the [6] approach, we succeeded the extraction of 4705 concepts out of which 2952 are relevant.

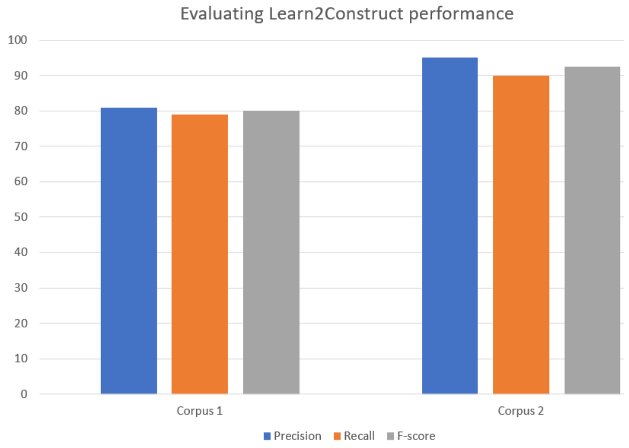


Figure 7: Evaluating Learn2Construct performance.

Table 3 shows the results of evaluating approaches for the concepts extraction using the Corpus 4. With the application of Learn2Construct approach, we succeeded the extraction of 60658 concepts of which 55689 are relevant according to DBpedia. On the other hand, when we adapted the approach in [6], we succeeded the extraction of 58652 concepts out of which 45958 are relevant.

	Extracted Concepts	Relevant Concepts
[Learn2Construct]	5145	4680
Anoop [6]	4705	2952

Table 2: Evaluating approaches for concepts extraction using corpus 3.

	Extracted Concepts	Relevant Concepts
[Learn2Construct]	60658	55689
Anoop [6]	58652	45958

Table 3: Evaluating approaches for concepts extraction using corpus 4.

	Extracted relations	Correctly identified relationships
[Learn2Construct]	2480	1879
	Concepts hierarchies	Validated Concepts hierarchies
Anoop [6]	1162	846

Table 4: Evaluating approaches for relations extraction using the corpus 3.

	Extracted relations	Correctly identified relationships
[Learn2Construct]	29562	23140
	Concepts hierarchies	Validated Concepts hierarchies
Anoop [6]	16524	12064

Table 5: Evaluating approaches for relations extraction using the corpus 4.

Regarding Tables 4 and 5, one can notice the superiority of Learn2Construct compared to [6] for the relations extraction using corpus 3 and corpus 4. This performance is explained by the auto-stabilization of LDA parameters (our approach adapts the optimal values of alpha, beta and the topics number by changing the corpus) for Learn2construct.
Concerning the relations extraction the performance of our approach can be explained by the use of linguistic patterns which is certainly more efficient than the concepts hierarchy used by [6].

5 CONCLUSION

In this paper, we have combined a machine learning approach which consists in the use of LDA in order to generate the document / topic clusters, as well as the topics / term clusters which suits the hundred documents which represent a corpus, with a linguistic approach consisting in the use of a lexico-syntactic-patterns in order to extract the semantic relations which connect our ontology concepts. We also evaluated the performance of our approach based on the precision, recall and F-score, which are the most recommended measures especially in the domain of information retrieval. We have automated the ontology construction process and even model stabilization in order to facilitate ontology reconstruction and its utilization with any application domain. In future work, we aim at extracting the hierarchical relations such as is-a, in order to sort the extracted concepts. Also, we would like to combine LDA with deep-learning algorithms to make our approach more efficient.

6 ACKNOWLEDGMENTS

This research work is supported by the Univ. Pau & Pays Adour, E2S UPPA, LIUPPA, OpenCEMS Industrial Chair.

REFERENCES

[1] Abeer Al-Arfaj and AbdulMalik Al-Salman. 2015. Arabic NLP Tools for Ontology Construction from Arabic Text: An Overview . *1st International Conference on Electrical and Information Technologies ICEIT'2015.* 6 (2015).

[2] Fatima N. Al-Aswadi, Huah Yong Chan, and Keng Hoon. 2019. Automatic ontology construction from text: a review from shallow to deep learning trend. *Springer* 28, 2 (November 2019), 5–6. https://link.springer.com/article/10.1007/s10462-019-09782-9

[3] Rubayyi Alghamdi and Khalid Alfalqi. 2015. A Survey of Topic Modeling in Text Mining. *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 1, 2015.* 7 (2015).

[4] Farman Ali, Daehan Kwak, Pervez Khan, Shaker El-Sappagh, Amjad Ali, Sana Ullah, Kye Hyun Kim, and Kyung-Sup Kwak. 2019. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *ELSEVIER* 17 (2019). https://www.sciencedirect.com/science/article/abs/pii/S0950705119300942

[5] David Alvarez-Melis and Martin Saveski. 2016. Topic Modeling in Twitter: Aggregating Tweets by Conversations. *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)* 4 (2016). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/viewFile/13162/12778

[6] V. S. Anoop, S. Asharaf, and P. Deepak. 2016. Unsupervised Concept Hierarchy Learning: A Topic Modeling Guided Approach. *Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016); ELSEVIER.* 9 (2016).

[7] Girija Attigeri, Manohara Pai M M, Radhika M Pai, and Rahul Kulkarni. 2018. Knowledge Base Ontology Building For Fraud Detection Using Topic Modeling. *ELSEVIER* 8 (2018).

[8] Kaveh Bastani, Hamed Namavari, and Jeffry Shaffer. 2019. Latent Dirichlet Allocation (LDA) for Topic Modeling of the CFPB Consumer Complaints. *Education and Management Engineering.* 34 (2019). https://www.sciencedirect.com/science/article/pii/S095741741930154X

[9] Stuart J. Blair, Yaxin Bi, and Maurice D. Mulvenna. 2019. Aggregated topic models for increasing social media topic coherence. *Springer* 19 (2019).

[10] Paul Buitelaar, Philipp Cimiano, and Bernado Magnini. 2005. Ontology Learning from Text: Methods, Evaluation and Applications. *Computational Linguistics, Volume 32, Number 4.* 4 (2005).

[11] S. Chowdhury and J. Zhu. 2019. Towards the Ontology Development for Smart Transportation Infrastructure Planning via Topic Modeling. *36th International Symposium on Automation and Robotics in Construction (ISARC 2019).* 10 (2019).

[12] Euthymios Drymonas, Kalliopi Zervanou, and Euripides G.M. Petrakis. [n.d.]. *ResearchGate* ([n. d.]).

[13] Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. 2020. Learning Interpretable and Statistically Significant Knowledge from Unlabeled Corpora of Social Text Messages: A Novel Methodology of Descriptive Text Mining. *Scitepress* 12 (2020).

[14] MINO George, P. BEAULAH SOUNDARABAI, and KARTHIK KRISHNAMURTHI. 2017. IMPACT OF TOPIC MODELLING METHODS AND TEXT CLASSIFICATION TECHNIQUES IN TEXT MINING: A SURVEY. *International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835* 6 (March 2017). http://www.iraj.in/journal/journal_file/journal_pdf/12-351-149622472172-77.pdf

[15] V Sree Harissh, M Vignesh, U Kodaikkaavirinaadan, and T V Geetha. 2017. Unsupervised Domain Ontology Learning From Text. *ResearchGate* 16 (2017).

[16] Pooja Kherwa and Poonam Bansal. 2018. Topic Modeling: A Comprehensive Review. *Researchgate* 17 (July 2018). https://www.researchgate.net/publication/334667298_Topic_Modeling_A_Comprehensive_Review

[17] Rinaldo Lima, Bernard Espinasse, and Fred Freitas. [n.d.]. *Engineering2019* ([n. d.]).

[18] Monika Rani, Amit Kumar Dhar, and O. P. Vyas. [n.d.]. Semi-Automatic Terminology Ontology Learning Based on Topic Modeling. *SEMANTIC SCHOLAR* 35 ([n. d.]). https://www.semanticscholar.org/paper/Semi-automatic-terminology-ontology-learning-based-Rani-Dhar/4948d5f16cd1f6733f2d989577119fdd18c83d02

[19] Satyaveer Singh and Mahendra Singh. 2019. ONTOLOGY LEARNING PROCEDURES BASED ON WEB MINING TECHNIQUES. *ResearchGate* 7 (2019).

[20] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring Topic Coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 952–961, Jeju Island, Korea, 12–14 July 2012..* 10 (2012).

[21] Weizhong Zhao, James J Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding, and Wen Zou. 2015. A heuristic approach to determine an appropriate number of topics in topic modeling. *Springer* 10 (2015). https://link.springer.com/article/10.1186/1471-2105-16-S13-S8

[22] Xiaofeng Zhu, Diego Klabjan, and Patrick N Bless. 2017. Unsupervised Terminological Ontology Learning based on Hierarchical Topic Modeling. *IEEE.* 10 (2017).