

Duplication-Driven Distributional Topic Modeling: A Catalyst for Strengthened Classification and Semantic Graphs

Amani Mechergui^{1,2,*}, Wahiba Ben Abdessalem Karaa^{1,2} and Sami Zghal³

¹High Institute of Management of Tunis, Tunis University, Tunisia

²RIADI Laboratory, National School of Computer Science, Manouba University, Tunisia

³University of Jendouba, FSJEGJ-LIPAH, Academic campus, Jendouba, Tunisia

Abstract

Topic modeling plays a crucial role in natural language processing and text mining by revealing underlying topic structures within textual data. This paper explores the integration of topic-seeding models with knowledge graphs to demonstrate their significance in shaping concepts and constructing semantic graphs. While the Latent Dirichlet Allocation (LDA) model serves as a foundational technique, the process of harmonizing textual data and KGs introduces challenges that necessitate innovative solutions. We propose the Duplication-Driven Distributional topic-seed-based LDA method that involves distributional term clustering over Functional Concepts (FCs), contributing to the creation and enrichment of a Universal Upper Semantic Graph (U2SG). It involves model fine-tuning, FC integration as seed terms, and the use of topic-seeding models. The key focus is on creating distinct, high-quality clusters by utilizing FCs and Noun Phrase patterns. Across diverse domains such as ontology and fishery, notable progress has been achieved. Our approach synergizes textual context with semantic nuances using techniques from both graph mining and machine learning, thereby augmenting the Knowledge Graph's comprehension. Our findings emphasize the effectiveness of this approach in constructing and enhancing the U2SG, showcasing its capacity to accommodate diverse concepts and adapt to specific domain-specific upper semantic graphs.

Keywords

Topic-seeding model, Latent Dirichlet Allocation, universal upper semantic graph, distributional term clustering, fundamental concepts, non-overlapping clusters, machine learning, noun phrase patterns, concept formation

1. Introduction

Topic modeling (TM) [1] is a prominent approach in the fields of natural language processing and text mining, playing an important role in uncovering the underlying topic structures inside large collections of textual data. This method involves the automatic recognition of topics or concepts that emerge from the complicated interplay of word occurrences within the text. The Latent Dirichlet Allocation (LDA) paradigm, proposed in 2003 by Blei, Ng, and Jordan [2], constitutes a foundational contribution to topic modeling, assuming documents are


TACC 2023: Tunisian-Algerian Joint Conference on Applied Computing, November 6 - 8, Sousse, Tunisia

*Corresponding author.

✉ amenimechergui47@gmail.com (A. Mechergui); wahiba.bak@gmail.com (W. Ben Abdessalem Karaa); zghal.sami@gmail.com (S. Zghal)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

compositions of latent topics defined by probabilistic distributions over words. Moreover, this methodology serves a dual purpose, functioning not only as a technique for clustering terms based on their distribution but also as a bridge between unstructured textual data and organized knowledge representation inherent in Knowledge Graphs (KGs). The KGs [3], repositories of interconnected entities, bestow valuable contextual insights into relationships, thus offering a rich layer that can significantly enhance the efficacy of TM. By furnishing supplementary information about term and concept semantics, KGs infuse depth into the TM process, elevating the comprehension of the underlying data.

The popularity of LDA stems from its generative nature and proficiency in capturing topic distribution across documents. However, a limitation arises as it autonomously attributes topic attributes to terms using a data-centric approach, disregarding prior knowledge during training. This oversight can culminate in clusters lacking semantic coherence and exhibiting overlapping tendencies.

The advent of topic-seed-based LDA models marks a significant advancement within the TM domain [4, 5]. This pioneering technique allows the direct infusion of preliminary topic cues into the model's training, a departure from conventional methods where topics arise solely from word co-occurrence patterns. Unlike the latter, topic-seed-based LDA models commence with predefined topic seeds, establishing a structured foundation for subsequent topic inference. By assimilating these seeds, the models can strategically concentrate on specific themes or concepts from the outset, fostering more precise and contextually fitting topic inferences. Consequently, the overall quality of the modeling process is enhanced. Nevertheless, their implementation alongside KGs presents a formidable challenge: effectively harmonizing KGs' semantic depth with textual data. These models often lean towards textual information over KG relationships, leading to an uneven portrayal of topics. This discrepancy may compromise the models' comprehensive grasp of the data's semantic nuances. In essence, their endeavor to reconcile textual and KG insights might impede the quality and pertinence of topic-seeding models. The integration of these models with KGs poses a critical challenge, as achieving a harmonious synthesis between KGs' semantic depth and textual data is pivotal.

In the upcoming sections of the paper, I will present a summary of the present status of my doctoral research, with a particular emphasis on the problem statement, motivation and research objectives, the methodologies employed, initial findings, and the broader significance of my work within the context of machine learning and its subdomain graph mining.

2. Problem Statement, Motivation, and Research Objectives

The application of topic-seeding models in conjunction with KGs reveals an inherent tension. These models often exhibit a bias towards relying on the textual information they are presented with, including the initial seed topics. In doing so, they may inadvertently overlook the structural intricacies embedded within the KG. This diversion of attention can lead to an imbalance where the influence of the pre-defined seed topics overshadows the potential impact of the intricate KG relationships. As a consequence, the model's capacity to holistically capture the subtle nuances of the data's semantics may be restricted, limiting its ability to represent the data in a truly comprehensive manner.

Hence, a central predicament arises when applying topic-seed-based LDA models in tandem with KGs: the intricate challenge of seamlessly and effectively fusing both the textual data and the inherent structural knowledge encapsulated within the KG. The models encounter difficulties in harmonizing these two distinct sources of information, potentially leading to a fragmented understanding of the data. As a result, the resultant topic representations might fall short of encapsulating the full semantic richness encoded within the dataset. This, in turn, could have repercussions on the overall quality and significance of the outcomes attained through this amalgamation.

Addressing this challenge is imperative due to the potential consequences it bears. The integration of topic-seeding models with KGs holds promise for richer insights by combining the textual context with KG's semantic depth. Failing to bridge this gap might result in inaccuracies and incomplete topic representations, limiting the models' ability to extract meaningful and relevant insights from complex datasets. Therefore, resolving this issue becomes essential for ensuring the robustness and applicability of topic-seeding models in real-world scenarios. The primary objective of my doctoral research is to address the challenge previously discussed by proposing an innovative approach that involves the creation of a catalyst designed to automate the classification of terms into Functional Concepts (FCs) classes through clustering. This not only aids in this classification process but also contributes to the development and enhancement of a Universal Upper Semantic Graph (U^2SG). The methodology employs a Duplication-Driven Distributional TM process. Furthermore, this approach involves the adaptation of topic seed-based LDA models, wherein FC-associated terms are selected as seed terms and topics are labeled with FCs. By leveraging the power of LDA and structured FC information, our research aims to fortify the term clustering framework and promote a deeper comprehension of semantic graphs. Motivations and Theoretical Foundations

3. Methodology

I have embraced a multi-faceted methodology to attain my research goals. Introducing "*the Duplication-Driven Distributional topic-seed-based LDA*" (D^3LDA), our method is centered on clustering terms related to FCs that shape Upper Semantic Graphs (USGs), forming a Universal Upper Semantic Graph (U^2SG). This involves constructing and enriching the U^2SG through Duplication-Driven Distributional Topic Modeling, guided by FCs present in our USGs.

Our research centers on exploring term clustering methods, notably focusing on topic seed-based LDA applications. Specifically, we concentrate on clustering terms aligned with FCs within a U^2SG . FCs, pivotal for defining other domain concepts, are of central interest in our investigation [6]. Additionally, we delve into FC classes, groups of terms referencing these FCs, encompassing synonyms, hyponyms, and related terms. This approach offers the advantage of narrowing the scope to synonym or hypernym relationships, optimizing computation. Our study proposes an inventive strategy for U^2SG construction and enhancement, achieved through a duplication-driven distributional TM process. Guided by USGs—models featuring FCs and their core interrelationships—this process defines the extensive U^2SG .

Our research's aim is to establish a holistic model for concept formation via term clustering. Leveraging LDA and structured FC information within a U^2SG , we enable efficient relationship

exploration and enrich the construction process. This approach is poised to advance concept relationship understanding and foster a refined representation of information in semantic graphs. In contrast to conventional topic-seeding models, our approach stands out for its ability to create non-overlapping clusters. This distinctiveness arises from the strict imposition of a constraint that prevents any overlap between two FC-seed sets. Furthermore, it incorporates FCs as initial seed terms, affording them predefined labels for individual topics, with the objective of clustering terms centered around the FCs associated with fishery, medicine, and ontology USGs. In Fig 1, we outline the interdependent key steps for term clustering over FCs using a topic-seeding model to construct and enhance a U^2SG from a generic corpus. Our proposed approach is composed of four respective Phases: (A) Data pretreatment, (B) 1st model training, (C) Bottom-up U^2SG constructing, and (D) 2nd model training.

To clarify, in the preliminary stage referred to as Phase A, our focal point entails the assembly of a pertinent and inclusive generic corpus encompassing scientific documents covering a diverse array of subjects, including fishery, ontology, and medicine. This diversity is integral to showcasing the significance of FCs within the corpus.

Our subsequent aim is the extraction of meaningful information from this corpus. To achieve this, a series of preprocessing steps are applied, encompassing the elimination of stop words, the removal of low-frequency words, conversion to lowercase and canonical form, and the application of part-of-speech filtering. This meticulous approach ensures the precision of outcomes.

Incorporating various natural language processing techniques, we emphasize the significance of Noun Phrase patterns due to their capacity to encapsulate intricate semantic insights. These patterns are pivotal for both concept extraction and SG construction, bolstering the overall quality of the endeavor.

Notably, the involvement of experts assumes a critical role in this process. Their expertise proves invaluable in identifying and rectifying any potential omissions during the initial cleansing phase. This extends to the detection and rectification of spelling errors and anomalies, ensuring the integrity and accuracy of the subsequent analysis. The outcome of this evaluation culminates in the formation of a set of potential candidate terms.

Hence, employing this collection of potential candidate terms as input during Phase B, our attention shifts towards fine-tuning the hyperparameters of our model and integrating pre-existing knowledge before initiating training. The goal is to generate topics that are not only interpretable but also closely aligned with our FCs. Following this, we will employ a multi-threshold technique on the model's topics to extract the most significant NPs, which will serve as the foundation for constructing a U^2SG . During Phase C, our primary goal is to create a U^2SG through a semi-automated and bottom-up approach. To achieve this, we integrate the topics, along with their associated subsumed and related NPs (bottom level) from Phase B. These NPs are manifested as descendants or leaves in the hierarchical tree. These elements are then progressively connected and organized into higher-level groupings, known as "upper" levels, more precisely the FCs of the USGs that collectively form the extensive U^2SG (*up level*). The resulting U^2SG comprises a total of 36 FCs, each adeptly characterizing our diverse USGs. To elaborate, within the ontology domain USG, there are 11 FCs, while the Fishery field USG encompasses 13 FCs. As for the medicine USG, we introduce a collection of 12 carefully selected FCs. This strategy offers numerous benefits, including semantic knowledge integration,

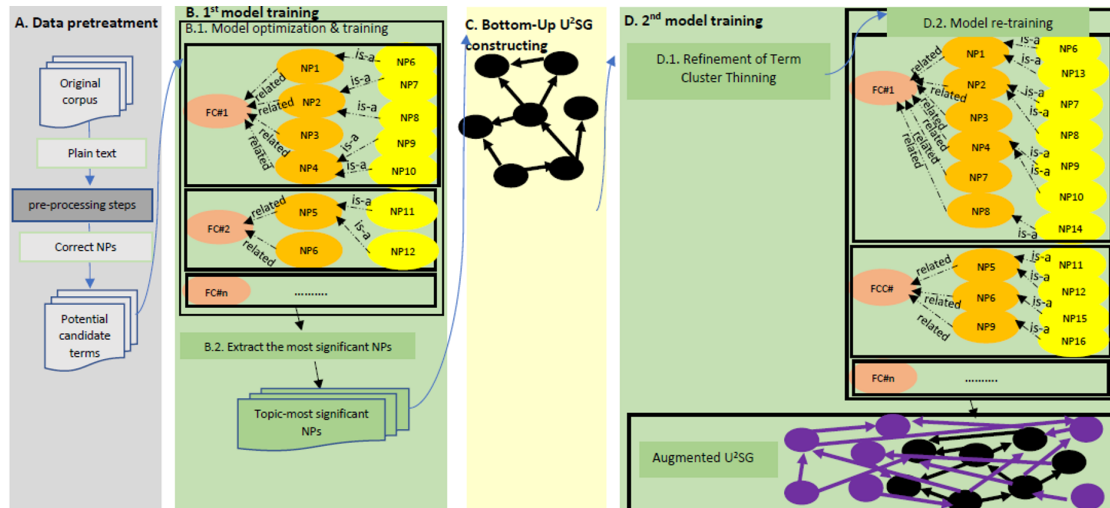


Figure 1: D^3LDA Steps for Term Clustering in a Topic Seeding Model over a U^2SG .

interoperability, scalability, reasoning, discovery, reusability, and sharing. It also features domain-agnosticism, and consistency, ultimately amplifying the effectiveness of semantic knowledge representation and engineering across diverse domains and artificial intelligence applications.

In Phase D, our central objective revolves around the further training of our model, focusing on the refinement of Term Cluster Thinning. Our aim remains to heighten the coherence and relevance of topic generation within our model. This entails the introduction of a pioneering constraint specifically for the new set of seed terms, added at this stage. Moreover, we amplify the impact of these selected seed terms by meticulously tuning the model's hyperparameters. This strategic Thinning of the term clustering process significantly elevates the quality and interpretability of topics, harmoniously aligning with the intricate context of our diverse data corpus. Following this, the topics that have been generated, along with their corresponding NPs, will be utilized to augment the previously constructed U^2SG during Phase C. This augmentation involves the incorporation of new NPs and the inclusion of newly identified relationships.

4. Current Progress

As of August 25, 2023, my research endeavors have yielded notable progress. Despite the ongoing nature of the study, I have obtained initial results that underscore the viability of my versatile proposed methodology across diverse domains such as ontology and fishery. This approach empowers me to craft outcome-focused topic representations that effectively encapsulate the intricate semantics embedded within the dataset. By combining textual context with the profound semantic dimensions inherent in the KG, this methodology not only provides me with deeper insights but also harmonizes the distinct sources of information. This harmonization results in a more precise alignment between textual data and the intrinsic structural knowledge enshrined within the KG. Consequently, the seamless integration of textual context and semantic nuances

enriches the overall understanding of the KG's content, fostering a holistic comprehension that is grounded in both the textual and structural aspects of the data.

The results underscore the potential of our proposed method to encompass a wide-ranging concept through term clustering, thereby presenting a unique opportunity to enhance the acquisition of USGs. This distinct advantage stems from the approach's adaptability across diverse domain-specific USGs, all while circumventing the need for additional annotation expenses. This flexibility broadens the horizons of effective TM application within the realm of USGs, capitalizing on the synergy between term clustering and domain-specific knowledge representation. As a result, our approach not only advances the efficiency of USG acquisition but also expands the applicability of TM, showcasing its potential as a versatile tool for uncovering meaningful insights within SKGs across various domains. I am delighted to share that I have achieved the successful publication of two noteworthy scientific papers. The first paper is titled "*A Bottom-Up Generic Probabilistic Building and Enriching Approach for Knowledge Graph using the LDA-Based Clustering Method*" [7]. This paper outlines a pioneering approach that employs LDA-based clustering to facilitate the bottom-up creation and enhancement of knowledge graphs. The second paper, titled "*Twice-Trained Agglomerative Clustering Approach using Topic Modeling over Generic Semantic Core Knowledge Graph*", [8], introduces an innovative clustering methodology. This approach involves utilizing TM in a twice-trained agglomerative clustering process, specifically designed to optimize the structure of a generic semantic core knowledge graph. Indeed, the outcomes presented in this second paper validate that our approach outperforms both semi-supervised and unsupervised distributional baselines on larger datasets compared to the first paper. This is vividly illustrated through the comparison, highlighting its notably superior performance. Both of these groundbreaking papers were presented at prominent conferences. The first paper was showcased at the "*21st IEEE/ACIS International Conference on Software Engineering, Management, and Applications (SERA 2023)*," an esteemed event that gathers experts and researchers in the fields of software engineering, management, and applications. The second paper found its place at the "*17th International Conference on Innovations in Intelligent Systems and Applications (INISTA 2023)*," a renowned conference dedicated to exploring the latest advancements and innovations in intelligent systems and their applications. These accomplishments stand as a testament to the tangible impact of my work. Through these publications, compelling evidence emerges, underscoring the innovativeness of my practical approach. This methodological framework has the inherent capacity to yield significant insights within the domain of machine learning. These results speak to the depth of my contributions, showcasing how my approach goes beyond conventional boundaries to unravel profound insights. At present, we are actively expanding our approach by incorporating a new domain, namely *the field of medicine*. This addition will bring the total number of study domains to three, allowing us to implement our approach and evaluate its versatility across various larger-scale and multilingual corpora. Elevating the commendable aspects of our proposition necessitates a comprehensive acknowledgment of its underlying limitations. Alongside its partial reliance on domain experts' existing knowledge, another substantial constraint lies in the potential sensitivity of the method to the quality and quantity of the initial labeled data. In scenarios where the initial dataset is sparse or contains inaccuracies, the efficacy of the approach might be compromised.

Moreover, the approach's adaptability to evolving domains and emerging lexicons should

be scrutinized. Rapid changes in domain-specific terminologies or the introduction of new concepts might necessitate frequent updates to the labeled data and potentially impact the approach's performance.

5. Conclusion and Future Outlook

Our study offers valuable insights into an innovative strategy to build and enhance a U^2SG using a duplication-based distributional approach for TM clustering along with LDA. Our primary focus centers on term clustering methodologies aimed at the formation of meaningful concepts. Our objective involves the adaptation of LDA to generate topics that align with the FCs of fish hunting and ontology-specific USGs, thus defining distinct modules within the overarching U^2SG framework. The findings from our two published papers validate that through extensive experimentation on two distinct datasets, our approach demonstrates its efficacy in constructing and subsequently enriching a U^2SG .

In the forthcoming times, our objective is to classify verbs and noun phrases together within a textual corpus. This involves gathering data in an additional language and subsequently training the model by incorporating it into the existing multilingual dataset. Additionally, the utilization of WordNet's existing hierarchy, coupled with the integration of supplementary prior knowledge from an expanded corpus, could prove to be of pivotal significance.

References

- [1] M. Apiola, M. Saqr, S. López-Pernas, The evolving themes of computing education research: Trends, topic models, and emerging research, *Past, Present and Future of Computing Education Research: A Global Perspective* (2023) 151–169.
- [2] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, *Journal of machine Learning research* (2003) 993–1022.
- [3] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: Opportunities and challenges, *Artificial Intelligence Review* (2023) 1–32.
- [4] K. Watanabe, A. Batur, Seeded sequential lda: A semi-supervised algorithm for topic-specific analysis of sentences, *Social Science Computer Review* (2023) 08944393231178605.
- [5] J. Wu, B. Li, Q. Liu, Topic detection based on bert and seed lda clustering model, in: *Proceedings of the 2023 7th International Conference on Innovation in Artificial Intelligence, 2023*, pp. 72–78.
- [6] L. Burita, P. Gardavsky, T. Vejlupek, K-gate ontology driven knowledge based system for decision support, *Journal of Systems Integration* 3 (2012) 19.
- [7] A. Mechergui, W. B. Abdessalem Karaa, S. Zghal, A bottom-up generic probabilistic building and enriching approach for knowledge graph using the lda-based clustering method, in: *21st IEEE/ACIS International Conference on Software Engineering, Management, and Applications (SERA 2023)*, 2023.
- [8] A. Mechergui, W. B. Abdessalem Karaa, S. Zghal, Twice-trained agglomerative clustering approach using topic modeling over generic semantic core knowledge graph, in: *2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA)*, 2023, pp. 1–6.