Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Prétraitement syntaxique pour enrichir le Bag of Words en Topic Modeling

Connor MacLean[1] Denis Cavallucci[1]
(1) INSA Strasbourg, 24 Bd de la Victoire, 67000 Strasbourg, France
connor.mac_lean@insa-strasbourg.fr, denis.cavallucci@insa-strasbourg.fr

RÉSUMÉ

Cet article propose une méthode de prétraitement innovante pour la *topic modeling* avec les modèles *Latent Dirichlet Allocation* (LDA) (Blei *et al.*, 2003) et *Embedding Topic Model* (ETM) (Dieng *et al.*, 2019), qui repose sur l'analyse des dépendances syntaxiques afin de construire des représentations plus riches du texte. En extrayant les têtes des groupes nominaux et verbaux ainsi que leurs compléments, notre approche génère des n-grammes syntaxiques (sn-grammes) plus informatifs que des bigrammes linéaires. Nous démontrons que cette stratégie permet de capturer les structures sémantiques complexes dans un corpus scientifique en français sur les énergies. Une évaluation expérimentale montre que, comparée à un prétraitement classique basé sur des unigrammes, notre approche accroît la diversité des sujets générés, tout en maintenant une cohérence raisonnable. Nous recommandons l'usage de métriques supplémentaires, telles que l'*Inversed Rank-Biased Overlap* (IRBO), pour évaluer cette diversité thématique. Nos résultats suggèrent que cette méthode enrichit la granularité des sujets extraits et permet des analyses plus fines de grands corpus textuels. Ce travail s'inscrit dans un projet de thèse de fouille de textes dans le but de mieux cibler des *startups* innovantes dans les énergies et les analyser selon la méthode TRIZ de résolution de contradictions techniques.

ABSTRACT

Syntactic Preprocessing to Enrich the Bag of Words in Topic Modeling

This article proposes an innovative preprocessing method for topic modeling with Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003) and Embedding Topic Models (ETM) (Dieng *et al.*, 2019), based on syntactic dependency analyses to build richer text representations. By extracting the heads of noun and verb phrases along with their complements, our approach generates syntactic n-grams (SN-grams) that are outperform bigrams. We demonstrate that this strategy captures complex structures in a corpus on energy. An experimental evaluation shows that, compared to a standard unigram-based model, our approach increases the diversity of generated topics while maintaining reasonable coherence. We recommend using additional metrics, such as Inversed Rank-Biased Overlap, to assess thematic diversity. Our results suggest that this method enhances the granularity of extracted topics and allows for more fine-grained analyses of large text corpora. This work is part of a thesis aimed at targeting innovative startups and analyzing them using the TRIZ method for resolving technical contradictions.

MOTS-CLÉS : LDA, ETM, dépendance syntaxique, SN-gramme, énergies renouvelables.

KEYWORDS: LDA, ETM, Syntactic Dependencies, SN-gram, renewable energy.

ARTICLE : **Accepté à** Atelier 4 AS - Atelier sur les Avancées en AMR et en Analyse Sémantiques.

1 Introduction

Le *topic modeling* (modélisation de sujets) est une technique d'apprentissage automatique non supervisée qui vise à identifier automatiquement les thématiques et à classifier de grands corpus textuels. Les méthodes modernes de *topic modeling* utilisent soit des approches statistiques et/ou probabilistes, soit s'appuient sur des représentations vectorielles de mots *word embeddings* et des modèles de langue existants afin d'extraire des informations à partir de texte non structuré.

Le *topic modeling* a connu des avancées significatives ces dernières années, avec des équipes explorant de nouvelles possibilités comme les modèles basés sur les *embeddings* (Dieng *et al.*, 2019), (Jiang *et al.*, 2020), (Terragni *et al.*, 2021), (Seifollahi *et al.*, 2021) et les méthodes de prétraitement basées sur les n-grammes (Zhu *et al.*, 2021), (Kherwa & Bansal, 2018), (Kawamae, 2014), (Nokel & Loukachevitch, 2016), à la fois pour les modèles à base d'*embeddings* que pour l'algorithme de *topic modeling* Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003).

Nous proposons une nouvelle façon, à base de sn-grammes (Sidorov *et al.*, 2013) de prétraiter des corpus en langue française pour améliorer l'utilité des modèles utilisant des corpus de type *Bag of Words (BoW)*. Nous constatons une amélioration de la diversité des mots trouvés au sein des *topics* sans dégradation de la cohérence.

2 Modèles

Le premier algorithme choisi pour illustrer cette technique de prétraitement est l'algorithme classique de *topic modeling* LDA (Blei *et al.*, 2003), sélectionné en raison de sa prédominance dans le domaine et servant de référence pour évaluer la performance de notre approche expérimentale.

Cet algorithme représente le corpus sous forme de sac de mots *Bag of Words*, c'est-à-dire que le vocabulaire est traité sans prise en compte de l'ordre des mots. Comme cette caractéristique est une limite de l'approche, de nombreux chercheurs ont proposé des méthodes visant à optimiser l'algorithme, notamment :

1. en ajoutant des mots voisins aux représentations des tokens (Zhu *et al.*, 2021)
2. en extrayant des expressions (Nokel & Loukachevitch, 2016)
3. en combinant des ontologies spécifiques à un domaine pour filtrer les corpus (Kim & Rhee, 2019)

Le second modèle utilisé est le modèle ETM (Dieng *et al.*, 2019) ou *embedding topic model*, qui rajoute des *word embeddings* au modèle tout en gardant l'approche *BoW*. Selon l'auteur, le modèle :

> ...modélise chaque mot à l'aide d'une distribution catégorielle dont le paramètre naturel est le produit scalaire entre une *embedding* du mot et une *embedding* du thème auquel il est assigné. (notre adaptation)

En combinant une approche moderne par *embeddings* avec un prétraitement en *BoW*, ce modèle permettra de montrer l'apport de notre nouvelle approche.

Le modèle CTM *(Contextualized Topic Model)*(Bianchi *et al.*, 2021b) a été considéré, mais nous ne l'utilisons pas, en raison du conseil de ne pas l'utiliser sur des corpus de plus de 200 documents.

3 SN-Grammes

Les *sn-grammes* sont un concept introduit pour la première fois par (Sidorov, 2013) en 2013. Deux applications distinctes ont été explorées par l'équipe : l'attribution d'auteur (Sidorov *et al.*, 2013) et la correction grammaticale pour l'anglais langue seconde (Sidorov, 2013). Ce concept diffère des techniques classiques de prétraitement n-grammes telles que les approches bigrammes, trigrammes et skipgrammes (Cheng *et al.*, 2006).

La principale différence entre les techniques n-grammes traditionnelles et les techniques sn-grammes réside dans la définition des éléments considérés comme voisins. Dans une approche n-grammes classique, les voisins sont simplement les *tokens* suivants et/ou précédents selon une fenêtre glissante. En revanche, dans une approche par sn-grammes, les voisins sont identifiés à l'aide d'un arbre de dépendance syntaxique.

En utilisant l'approche par sn-grammes, nous pouvons relier les noms et leurs compléments en analysant l'arbre syntaxique et en identifiant les dépendances correspondantes. Par exemple, il est possible de former des n-grammes à partir d'adjectifs et du nom tête du syntagme. Des n-grammes peuvent également être formés à partir du verbe et de son complément, afin de prendre aussi en compte les actions.

Grâce à l'association par voisins syntaxiques, il devient possible de créer des n-grammes sémantiquement riches, tels que *bien public* et *énergie renouvelable*, même si les mots individuels se trouvent à des distances de plusieurs tokens. Par exemple, dans la phrase :

"La transition vers une économie bas carbone nécessite une politique ambitieuse en matière énergétique."

Les bigrammes simples linéaires détecteront des séquences comme :

1. transition vers
2. économie bas
3. politique ambitieuse
4. matière énergétique

Mais ils manqueront l'association cruciale entre "politique" et "énergétique", qui sont grammaticalement liés mais séparés par des mots intermédiaires.

Une approche fondée sur les dépendances syntaxiques permet, elle, d'identifier "politique énergétique" comme une unité conceptuelle, car elle suit une relation de modification du nom par un adjectif postposé — fréquente en français.

Ces associations syntaxiques révèlent donc des collocations thématiquement riches, impossibles à détecter par des méthodes à base de n-grammes classiques.

4 Nouveau traitement canonique des groupes nominaux (GN) et verbaux (GV)

4.1 Contexte

Notre approche novatrice s'inspire de travaux antérieurs ayant recours à l'analyse syntaxique en dépendances pour des tâches de traitement en aval, tels que (Gamallo *et al.*, 2012), (Delpisheh & An, 2014), et (Wang *et al.*, 2007). Bien que notre méthode soit similaire à l'approche des SN-grammes présentée par (Sidorov, 2013), nous nous concentrons exclusivement sur les noms associés à leur(s) complément(s) ainsi que sur les verbes et leur(s) complément(s).

À ce jour, aucun consensus n'existe quant à la meilleure méthode pour extraire des bigrammes, trigrammes ou n-grammes à partir de corpus tout en conservant un maximum de contexte et d'informations pertinentes. Plusieurs équipes ont proposé leurs solutions, notamment (Nokel & Loukachevitch, 2016), (Zhu *et al.*, 2021), (Almgerbi *et al.*, 2021), et (Kawamae, 2014). Ces travaux soulignent à la fois la nécessité d'une standardisation des approches de *topic modeling* par n-grammes, et les avancées significatives réalisées lorsque les n-grammes sont intégrés dans les étapes de prétraitement.

4.2 Méthodologie

Notre approche vise à créer des n-grammes basés sur les voisins syntaxiques de l'arbre de dépendances. Concrètement, l'approche consiste à :

1. Trouver les noms et les verbes, en faire des unigrammes
2. Créer des n-grammes à partir de noms et de verbes et leurs compléments, par exemple "énergie renouvelable verte" == énergie_renouvelable & énergie_verte & énergie_renouvelable_verte

Pour arriver à repérer les dépendances syntaxiques, nous utilisons la librairie spaCy (Honnibal & Montani, 2017) et son modèle "fr_news_core_lg", ce qui permet un traitement rapide et suffisamment précis pour le prototypage du processus. Depuis son intégration avec HuggingFace [1], il est possible de remplacer le modèle de repérage de dépendances syntaxiques par un modèle disponible sur HuggingFace. Cela permettra de mettre à jour le modèle au besoin. Nous utilisons les structures de type *noun chunk* intégrée dans la librairie. Ensuite, nous codons manuellement une fonction permettant le traitement des groupes verbaux *verb chunks*. Nous évitions l'utilisation des IA génératives en raison de la durée de traitement des corpus de grande taille et de l'ajout possible des erreurs dans la sortie. Dans notre processus, il n'est également pas nécessaire de supprimer les *stopwords*, car notre approche ne les prend pas en compte.

Pour les premières expériences, nous avons d'abord utilisé un corpus de 462 résumés d'articles scientifiques du domaine des énergies renouvelables extrait d'Istex [2]. Malheureusement, après un filtrage par langdetect [3], il ne restait plus que 210 articles reconnus comme étant en français. Nous avons donc décidé d'élargir notre recherche et de créer un corpus sur les énergies et non spécifiquement les énergies renouvelables. Nous avons donc, après le filtrage initial, 7 294 résumés d'articles scientifiques. Après le prétraitement, nous montrons la différence de taille de corpus (en tokens) selon trois méthodes de prétraitement. Concrètement :

1. https ://huggingface.co/
2. https ://www.istex.fr/
3. https ://pypi.org/project/langdetect/

— 322 405 : Longueur du corpus unigramme sans stopwords
— 237 276 : Longueur du corpus SN-gramme
— 843 941 : Longueur du corpus en bigrammes simples

Pour notre analyse, nous avons entraîné deux modèles de type LDA et deux modèles de type ETM (Dieng *et al.*, 2019). Le premier est entraîné sur notre corpus *BoW* sans *stopwords* et est constitué d'unigrammes. Le deuxième est entraîné sur notre corpus de sn-grammes. Nous avons prétraité le corpus en créant des unigrammes sans stopwords pour le premier modèle, et nous avons prétraité le corpus selon notre nouvelle méthode pour le second.

Une évaluation de cette technique sera menée sur des corpus de grande taille en anglais dans la suite de ce projet.

4.3 Évaluation

Notre nouvelle approche parvient à démontrer son utilité en termes de diversité thématique. Pour notre première expérience, nous avons décidé de générer 9 *topics* par modèle et 25 termes générés par *topic*. Nous commençons par comparer la proportion de termes uniques générés dans les 9 *topics* générés par nos modèles. Si un terme est présent dans les 25 premiers mots de plusieurs *topics*, ou plusieurs fois dans le 25 premiers termes d'un seul *topic*, la mesure de diversité diminue. Un score parfait de 100% signifierait que tous les *topics* (25 termes × 9 *topics*) sont remplis par des termes uniques, sans aucune répétition entre ou au sein des *topics*. Les résultats de cette analyse sont présentés dans le Tableau 1.

Pour obtenir une analyse plus robuste de la diversité thématique entre nos deux modèles, nous utilisons une deuxième métrique, présentée dans (Bianchi *et al.*, 2021a), la diversité thématique via l'*Inversed Rank-Biased Overlap (IRBO)* :

> L'Inversed Rank-Biased Overlap (p) mesure la diversité des topics générés par un modèle. Il s'agit de l'inverse du RBO standard (Webber et al., 2010 ; Terragni et al., 2021b), qui compare les 10 mots les plus représentatifs de deux topics en prenant en compte à la fois les mots partagés et leur position dans la liste. Ainsi, des mots identiques à des rangs différents sont moins pénalisés que ceux en première position. La valeur de p est de 0 pour des topics identiques et de 1 pour des topics complètement différents. (notre adaptation)

Contrairement à des mesures comme la diversité lexicale brute, l'IRBO pénalise moins les répétitions de mots situés à des rangs éloignés, reflétant mieux la diversité sémantique

Nous avons également questionné la pertinence des méthodes d'évaluation automatique telles que la cohérence thématique, dont les limites sont bien mises en évidence par (Fan *et al.*, 2019) :

> Les métriques classiques de qualité des topics ne sont pas robustes aux mots vides (*stopwords*). Nous montrons que deux mesures standards — la cohérence et le PMI — réagissent de manière contre-intuitive lorsque le corpus contient de nombreux mots fréquents mais peu informatifs. Cette situation est fréquente dans les corpus réels, où un vocabulaire standardisé est souvent répété dans le texte, mais reste peu informatif. Cela dit, notre analyse ne remet pas en cause l'usage de ces métriques dans les cas où le vocabulaire a été rigoureusement filtré pour sa pertinence. Après avoir discuté

de la cohérence et du PMI, nous introduisons une autre mesure, le log lift, qui permet d'atténuer les problèmes liés aux mots vides. (notre traduction)

Nous analysons donc une mesure traditionnelle de cohérence en parallèle avec les mesures de diversité pour nos modèles. Nous utilisons la mesure de cohérence c_uci, proposée initialement par (Newman *et al.*, 2010) et implémentée via la bibliothèque Gensim (Řehůřek & Sojka, 2010). Cette mesure permet de qualifier les résultats d'un modèle de "cohérent" en fonction de son score. Un score plus haut signifie que les mots générés par le modèle sont sémantiquement plus proches et donc plus pertinents. Un score bas, voire négatif, signifie que les mots générés par le modèle ne sont pas proches et que les *topics* peuvent contenir plus de bruit. Les résultats de cette comparaison sont présentés dans le Tableau 1.

Modèle	**Cohérence**	**Mots Uniques/Topic (%)**	**IRBO**
(LDA) Modèle Unigramme	-1.2	56.00	0.77
(LDA) Modèle SN-gramme	-2.4	62.22	0.87
(ETM) Modèle unigramme	-0.1	49.33	0.75
(ETM) Modèle SN-gramme	0.1	77.78	0.97

TABLE 1 – Cohérence vs. Diversité entre les approches de *topic modeling*

Comme le montre le Tableau 1, en comparant notre approche avec un modèle unigramme classique, on observe une nette augmentation de la diversité thématique. Nous constatons également une nette amélioration de la cohérence du modèle quand un modèle de type ETM est utilisé.

5 Discussion & Suite du projet

Notre méthode démontre trois avantages principaux par rapport aux approches traditionnelles :

1. une extraction de termes plus pertinents grâce au contexte syntaxique
2. une réduction significative de la taille du corpus comparé aux méthodes bigrammes classiques/simples
3. une amélioration notable de la diversité thématique

L'analyse révèle particulièrement que les modèles ETM bénéficient davantage de cette approche, affichant des performances supérieures en diversité et cohérence thématique par rapport aux implémentations LDA.

Bien que les résultats soient prometteurs, cette étude présente certaines limitations. La taille restreinte de notre corpus expérimental (7,294 articles) ne permet pas d'évaluer pleinement la scalabilité de la méthode. Les travaux futurs s'orienteront vers trois axes principaux :

1. L'évaluation sur des corpus étendus, incluant aussi bien des résumés que des articles complets
2. L'adaptation de la méthode à des corpus multilingues, particulièrement en anglais
3. Des tests de scalabilité sur des corpus de plusieurs millions de documents

Ces développements permettront de mieux cerner le potentiel de notre approche pour le traitement de corpus de grande taille.

Remerciements

Ce travail a été réalisé au sein de l'équipe CSIP du laboratoire ICube, à l'INSA de Strasbourg, dans le cadre d'un contrat doctoral. Nous tenons à exprimer notre gratitude à tous nos collègues pour leur soutien, leur collaboration et les idées qu'ils ont apportées à ce projet.

Nous adressons une reconnaissance particulière au groupe Discovery d'EDF (Électricité de France) pour leur collaboration inestimable tout au long de ce projet. Leur soutien et leurs perspectives ont enrichi les résultats de cette recherche.

Références

ALMGERBI M., DE MAURO A., KAHLAWI A. & POGGIONI V. (2021). Improving Topic Modeling Performance through N-gram Removal. p. 162–169. DOI : 10.1145/3486622.3493952.

BIANCHI F., TERRAGNI S. & HOVY D. (2021a). *Pre-training is a Hot Topic : Contextualized Document Embeddings Improve Topic Coherence*. Pages : 766, DOI : 10.18653/v1/2021.acl-short.96.

BIANCHI F., TERRAGNI S., HOVY D., NOZZA D. & FERSINI E. (2021b). Cross-lingual Contextualized Topic Models with Zero-shot Learning. In P. MERLO, J. TIEDEMANN & R. TSARFATY, Éds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 1676–1683, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.143.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**(null), 993–1022.

CHENG W., GREAVES C. & WARREN M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, **11**(4), 411–433. Publisher : John Benjamins, DOI : 10.1075/ijcl.11.4.04che.

DELPISHEH E. & AN A. (2014). Topic Modeling Using Collapsed Typed Dependency Relations. In V. S. TSENG, T. B. HO, Z.-H. ZHOU, A. L. P. CHEN & H.-Y. KAO, Éds., *Advances in Knowledge Discovery and Data Mining*, p. 146–161, Cham : Springer International Publishing. DOI : 10.1007/978-3-319-06605-9_13.

DIENG A. B., RUIZ F. J. R. & BLEI D. M. (2019). Topic Modeling in Embedding Spaces.

FAN A., DOSHI-VELEZ F. & MIRATRIX L. (2019). Assessing topic model relevance : Evaluation and informative priors. *Statistical Analysis and Data Mining : The ASA Data Science Journal*, **12**(3), 210–222. Publisher : Wiley Online Library.

GAMALLO P., GARCIA M. & FERNÁNDEZ-LANZA S. (2012). Dependency-Based Open Information Extraction. In O. ABEND, C. BIEMANN, A. KORHONEN, A. RAPPOPORT, R. REICHART & A. SØGAARD, Éds., *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, p. 10–18, Avignon, France : Association for Computational Linguistics.

HONNIBAL M. & MONTANI I. (2017). spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

JIANG Z., SRIVASTAVA M., KRISHNA S., AKODES D. & SCHWARTZ R. (2020). Combining Word Embeddings and N-grams for Unsupervised Document Summarization. arXiv :2004.14119 [cs, stat], DOI : 10.48550/arXiv.2004.14119.

KAWAMAE N. (2014). Supervised N-gram topic model. *Proceedings of the 7th ACM international conference on Web search and data mining*, p. 473–482. Conference Name : WSDM 2014 : Seventh

ACM International Conference on Web Search and Data Mining ISBN : 9781450323512 Place : New York New York USA Publisher : ACM, DOI : 10.1145/2556195.2559895.

KHERWA P. & BANSAL P. (2018). Semantic N-Gram Topic Modeling. *ICST Transactions on Scalable Information Systems*, **0**(0), 163131. DOI : 10.4108/eai.13-7-2018.163131.

KIM H. H. & RHEE H. (2019). An Ontology-Based Labeling of Influential Topics Using Topic Network Analysis. *Journal of Information Processing Systems*, **15**, 1096–1107.

NEWMAN D., LAU J. H., GRIESER K. & BALDWIN T. (2010). Automatic Evaluation of Topic Coherence. In R. KAPLAN, J. BURSTEIN, M. HARPER & G. PENN, Éds., *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p. 100–108, Los Angeles, California : Association for Computational Linguistics.

NOKEL M. & LOUKACHEVITCH N. (2016). Accounting ngrams and multi-word terms can improve topic models. In *Proceedings of the 12th Workshop on Multiword Expressions*, p. 44–49, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/W16-1806.

SEIFOLLAHI S., PICCARDI M. & JOLFAEI A. (2021). An Embedding-Based Topic Model for Document Classification. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, **20**(3), 52 :1–52 :13. DOI : 10.1145/3431728.

SIDOROV G. (2013). Syntactic dependency based n-grams in rule based automatic English as second language grammar correction. *Int. J. Comput. Linguistics Appl.*, **4**(2), 169–188.

SIDOROV G., VELASQUEZ F., STAMATATOS E., GELBUKH A. & CHANONA-HERNÁNDEZ L. (2013). Syntactic Dependency-Based N-grams as Classification Features. In I. BATYRSHIN & M. G. MENDOZA, Éds., *Advances in Computational Intelligence*, Lecture Notes in Computer Science, p. 1–11, Berlin, Heidelberg : Springer. DOI : 10.1007/978-3-642-37798-3_1.

TERRAGNI S., FERSINI E. & MESSINA V. (2021). Word Embedding-Based Topic Similarity Measures. p. 33–45. DOI : 10.1007/978-3-030-80599-9_4.

WANG X., MCCALLUM A. & WEI X. (2007). Topical N-Grams : Phrase and Topic Discovery, with an Application to Information Retrieval. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, p. 697–702. ISSN : 2374-8486, DOI : 10.1109/ICDM.2007.86.

ZHU L., HUANG M., CHEN M. & WANG W. (2021). An N-gram based approach to auto-extracting topics from research articles. *ArXiv*.

ŘEHŮŘEK R. & SOJKA P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*, p. 45–50, Valletta, Malta : ELRA.