Utiliser les connaissances du sens commun pour la découverte des topics intérpretables

I. Harrando[1], R. Troncy[1]

[1] EURECOM, Sophia Antipolis

mél

Résumé

Les approches traditionnelles de modélisation de sujets (Topic Modeling) s'appuient généralement sur des statistiques de cooccurrence entre termes et documents pour trouver des sujets latents dans une collection de documents. Cependant, le fait de s'appuyer uniquement sur ces statistiques peut donner des résultats incohérents ou difficiles à interpréter pour les utilisateurs finaux dans de nombreuses applications où l'intérêt réside dans l'interprétation des sujets résultants (e.g. l'étiquetage de documents, la comparaison de corpus, orienter l'exploration du contenu..). Nous proposons de tirer parti des connaissances externes de sens commun, c'est-à-dire des informations du monde réel au-delà de la cooccurrence des mots, pour trouver des topics plus cohérents et plus facilement interprétables par les humains. Nous présentons le "Common Sense Topic Model" (CSTM), une approche nouvelle et efficace qui augmente le clustering avec des connaissances extraites du graphe de connaissances ConceptNet. Nous évaluons cette approche sur plusieurs jeux de données en comparaison avec des modèles couramment utilisés, en utilisant une évaluation automatique et humaine, et nous montrons comment elle montre une corrélation supérieure au jugement humain.
Cet article a été déjà publié à K-CAP 2021[4].

Mots-clés

Exemple type, format, modèle.

1 Introduction

Le Topic modeling (modélisation de sujets) est une technique de fouille de textes qui est largement utilisée pour de nombreuses applications, à la fois pour d'autres tâches dites "downstream" du TAL (e.g. la similarité de textes), mais aussi comme un outil pour explorer, visualiser et interpréter le contenu de grandes collections de textes. Alors que la première application peut être évaluée et améliorée en mesurant quantitativement la performance sur la tâche elle-même, il est plus difficile de saisir la capacité d'un algorithme de topic modeling à générer des résultats compréhensibles et utiles pour un utilisateur humain. Plusieurs efforts de recherche antérieurs [1, 2] ont mis en évidence la divergence entre la plupart des mesures d'évaluation automatiques (largement utilisées dans la littérature) et le jugement humain, car ces modèles ont tendance à optimiser pour des objectifs numériques qui s'alignent ou se corrèlent rarement bien avec ce que les humains considèrent comme des "sujets" (topics).

La plupart des approches de modélisation des sujets se concentrent sur les statistiques de co-occurrence des mots comme signal principal pour détecter les relations sémantiques latentes entre eux – une idée qui remonte aux années 50 (*"Vous connaîtrez un mot par la compagnie qu'il garde"*[3]). Cela les rend intrinsèquement incapables de capturer les relations entre les mots qui ne sont pas explicitement présents dans les données d'apprentissage. De nombreux travaux ont été réalisés pour explorer la possibilité d'injecter des connaissances externes (généralement spécifiques à un domaine) dans la tâche de modélisation de sujets. Pourtant, bien que l'utilisation du sens commun a été explorée pour la classification des topics [5], aucune tentative d'incorporation de connaissances générales humaines (ou *sens commun*) dans le processus de modélisation de sujets n'a été proposée pour combler le fossé entre l'optimisation basée sur les statistiques et le jugement humain. Nous proposons une méthode qui combine les connaissances dans un graphe de connaissances dit de sens commun [7] avec un algorithme de clustering pour produire des sujets qui sont plus corrélés avec le jugement humain de la cohérence tout en s'adaptant sans problème à de grands ensembles de données.

2 Approche

Comme dans les travaux précédents [6], nous abordons la tâche de modélisation des sujets comme un *problème de clustering de documents*, c'est-à-dire que nous générons des représentations vectorielles pour tous les documents du corpus étudié que nous appelons *Sac de mots enrichi en sens commun (Common-sense enriched Bag-of-words)*, puis nous exécutons un algorithme de clustering pour trouver N groupes cohérents (N étant le nombre de sujets) qui représentent nos sujets. On appelle ce modèle *CSTM* (Common-Sense Topic Model). La figure 1 représente illuste cette approche.

Common-sense Enriched Bag of Words (CS-BoW). Inspirés par des méthodes issues de la littérature sur l'expansion de requêtes, nous proposons d'enrichir la représentation "sac de mots" des documents, avec des termes connexes issus du graphe de connaissances ConceptNet.

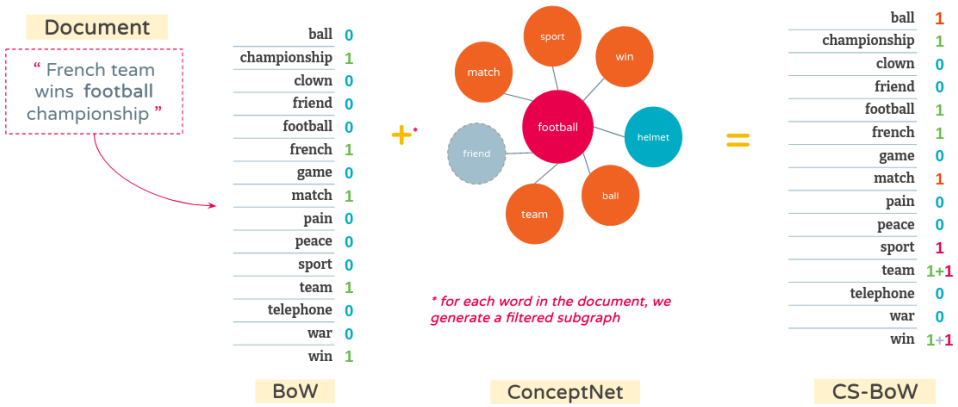

FIGURE 1 – Illustration of the process of creating the common sense-enhanced document representation using ConceptNet. We note that we filter out words that do no appear in the vocabulary (friend) as well as words with low similarity (helmet).

L'avantage d'utiliser ConceptNet est qu'il est principalement peuplé par la relation "Relié à", qui implique une relation topique entre les termes. Concrètement, pour chaque mot du document, nous interrogeons ConceptNet pour récupérer tous les termes qui lui sont directement liés (à un saut de puce sur le graphe), et nous les ajoutons au document. Par exemple, un document qui mentionne le mot "caméra" sera automatiquement enrichi avec les mots "photo", "objectif", etc. La représentation du document est alors construite comme un sac de mots contenant tous les mots originaux du document, en plus de tous les mots qui leur sont liés dans ConceptNet.

Clustering. Il existe une littérature riche et variée sur la tâche de clustering. Dans un souci de simplicité, nous choisissons *K-Means*, un algorithme de clustering couramment utilisé, rapide et capable de traiter des ensembles de données plus importants à l'aide de l'implémentation hautement optimisée *FAISS* [1], et nous l'exécutons sur les représentations CS-BoW des documents du corpus. Pour générer les mots clés du sujet, nous considérons les vecteurs centroïdes générés par K-Means et choisissons les composantes (correspondant aux mots sur la représentation CS-BoW) avec les plus grands coefficients pour représenter le sujet.

3 Evaluation

On propose de faire l'évaluation de notre modèle (en le comparant avec deux autres baselines) en deux étapes : évaluation automatique (quantitative) avec les métriques classiques utilisées dans la litérature, et puis une évaluation qualitative faite par 12 personnes qui parlent l'anglais couramment. On leur demande d'effectuer trois tâches pour évaluer les topics résultants (Intrusion de mots, labélisation des sujets, classification des sujets).

On observe globalement que malgré le fait que CSTM n'est pas toujours le meilleur au niveau des métriques automatiques, il dépasse de manière considérablement les autres modèles sur les tâches de l'évaluation humaine, ce qui montre que les sujets générés par CSTM sont plus facilement intérprétés par les humains.

1. https://github.com/facebookresearch/faiss/

Remerciements

Ce travail a été partiellement soutenu par le programme de recherche et d'innovation Horizon 2020 de l'Union européenne dans le cadre du projet Odeuropa (accord de subvention n° 101004469), par CHIST-ERA dans le cadre du projet CIMPLE (CHIST-ERA-19-XAI-003) et par raisin.ai dans le cadre du projet MyLittleEngine.

Références

[1] Jonathan CHANG et al. “Reading Tea Leaves : How Humans Interpret Topic Models”. In : NIPS'09. Vancouver, Canada, 2009.

[2] Caitlin DOOGAN et Wray BUNTINE. “Topic Model or Topic Twaddle ? Re-evaluating Semantic Interpretability Measures”. In : *NAACL '21*. Juin 2021.

[3] Adriana FERRUGENTO et al. *A synopsis of linguistic theory 1930-1955*. 1957.

[4] Ismail HARRANDO et Raphaël TRONCY. “Discovering Interpretable Topics by Leveraging Common Sense Knowledge”. In : K-CAP '21. USA, 2021.

[5] Ismail HARRANDO et Raphaël TRONCY. “Explainable Zero-Shot Topic Extraction Using a Common-Sense Knowledge Graph”. In : *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Dagstuhl, Germany.

[6] Suzanna SIA, Ayush DALMIA et Sabrina J. MIELKE. “Tired of Topic Models ? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too !” In : *EMNLP '20*. Online : ACL, nov. 2020.

[7] Robyn SPEER, Joshua CHIN et Catherine HAVASI. “ConceptNet 5.5 : An Open Multilingual Graph of General Knowledge”. In : 2017, p. 4444-4451.