REDIRE : Réduction Extrême de DImension pour le Résumé Extractif

Marius Ortega*, Aurélien Bossard,** Nédra Mellouli,*,** Christophe Rodrigues*

*Léonard De Vinci Pôle Universitaire, Research Center, 92916 Paris La Défense, France
**Laboratoire d'Intelligence Artificielle et Sémantique des Données,
Université Paris 8 (EA4383), 93200 Saint-Denis, France
marius.ortega@edu.devinci.fr; nedra.mellouli@devinci.fr;
aurelien.bossard@iut.univ-paris8.fr; christophe.rodrigues@devinci.fr

Résumé. Nous présentons un modèle de résumé automatique non supervisé capable d'extraire les phrases les plus importantes d'un ensemble de textes. Pour extraire les phrases dans un résumé, nous utilisons des plongements de mots pré-entraînés afin de représenter les documents. A partir de cet épais nuage de vecteurs de mots, nous appliquons une réduction extrême de dimension permettant d'identifier des mots importants que nous regroupons par proximité. Les phrases sont extraites grâce à l'optimisation linéaire pour maximiser l'information présente dans le résumé. Nous évaluons l'approche sur des documents de grande taille et présentons des premiers résultats très encourageants.

1 Introduction

Dans cet article, nous présentons une méthode de Réduction Extrême de DImension pour le Résumé Extractif (REDIRE). Cette méthode est entièrement non supervisée. Si les méthodes génératives supervisées ont vu le plus de progrès notables depuis *Pointer-Generator* See et al. (2017), le premier système de résumé génératif par apprentissage profond, les méthodes extractives restent nécessaires pour s'affranchir des corpus d'apprentissage imposants et coûteux, des limites de domaine et des limites liées à la taille des documents en entrée.

L'idée de cette méthode nous est venue d'une étude sur les plongements de mots, que nous présentons en Section 3. Nous décrivons la méthode en Section 4 avant de montrer en Section 5 qu'elle s'applique avec de bons résultats à des documents longs.

2 Travaux connexes

TextRank Mihalcea et Tarau (2004) est sûrement la *baseline* la plus utilisée en résumé automatique non supervisé. Elle sélectionne les phrases des documents source d'après leur centralité dans une représentation graphique du document fondée sur les similarités entre phrases, grâce à un algorithme de marche aléatoire dans le graphe. Gillick et Favre (2009) considèrent le résumé comme un problème d'optimisation de sac à dos visant à maximiser la somme des

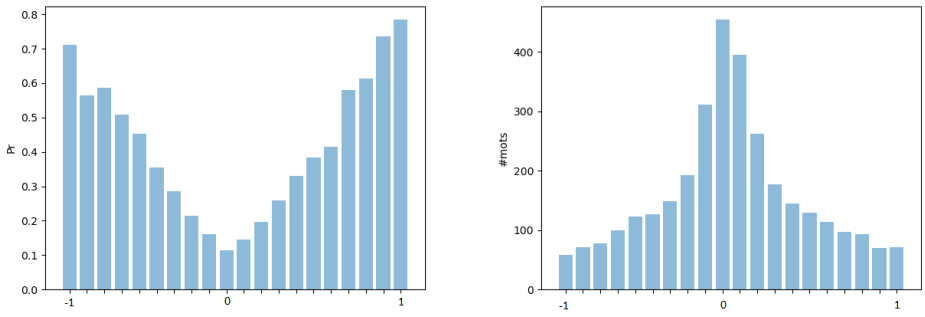

FIG. 1 – (gauche) : Proportion de mots appartenant aux résumés de référence dans des intervalles de distance des mots (projetés en une dimension) à l'origine. (droite) : Dispersion des mots de l'ensemble des documents sur l'axe sur lequel ils ont été projetés.

fréquences des mots, calculée sur les documents source, au sein du résumé, et ceci en respectant une contrainte de taille du résumé en nombres de mots. Pour cela, ils utilisent une méthode de programmation linéaire en nombres entiers. PMI (Padmakumar et He, 2021) repose sur un modèle de langage utilisé pour estimer la probabilité d'un document étant donné une requête, à la manière du modèle de vraisemblance de Manning et al. (2008). La requête est ici remplacée par une phrase candidate à l'extraction. Un algorithme glouton extrait les phrases en fonction de l'estimation de probabilité calculée. SummPip (Zhao et al., 2020) est une méthode de résumé multi-documents non supervisée qui vise à générer un graphe des documents à résumer dans lequel les nœuds sont les phrases et les arêtes construites grâce à des indices de surface, des informations sémantiques exogènes et des similarités sémantiques. Le graphe est alors partitionné selon un algorithme du plus court chemin avant la sélection d'une phrase. SummVD (Shenouda et al., 2022) utilise la décomposition en valeurs singulières afin de réduire la dimensionnalité des plongements de mots, chaque nouvelle dimension représentant alors un sujet latent. Les phrases porteuses des mots les plus proches des sujets latents identifiés sont retenues dans le résumé.

3 Étude préliminaire

La grande dimension des plongements de mots les rend difficiles à expliquer et à interpréter. Beaucoup de travaux vont dans ce sens (Mimno et Thompson, 2017). Nous nous intéressons à la détection de zones de l'espace qui pourraient potentiellement concentrer les mots spécifiques d'un texte. C'est dans cet objectif que nous avons décidé de réduire à l'extrême la dimension des plongements de mots à l'aide de la méthode UMAP (McInnes et al., 2020), une méthode de réduction de dimension fondée sur la géométrie riemanienne et l'algèbre topologique. Après avoir réduit à l'extrême, à une dimension, des plongements de mots issus d'un document, nous avons constaté que les mots les plus éloignés de l'axe semblaient également les plus spécifiques, donc les plus à même de caractériser le document. Nous proposons de valider cette hypothèse sur un corpus de résumé automatique présenté en §5.1.

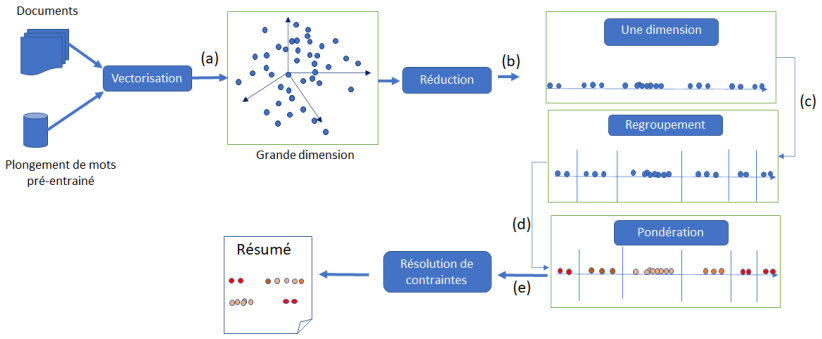

FIG. 2 – Workflow de notre approche non-supervisée : construction de résumés extractifs

Les résumés de référence du corpus peuvent servir à évaluer le rapport entre la position d'un mot dans l'espace et sa spécificité. Nous avons réduit la dimension des plongements de mots calculés d'après FastText (Mikolov et al., 2018), de 300 à 1 dimension. Nous avons alors réalisé une étude visant à infirmer ou confirmer l'hypothèse selon laquelle la position d'un mot dans l'espace en 1 dimension est corrélée à sa spécificité ou son importance dans le document à résumer.

La figure 1 (gauche) présente pour chaque intervalle de distance des plongements de mots projetés en 1 dimension à l'origine de l'axe, la proportion des mots des documents également présents dans le résumé de référence. On constate que plus on s'éloigne de l'origine, plus les mots sont susceptibles d'appartenir au résumé de référence. La figure 1(droite) présente la position des mots qui appartiennent aux documents source. On constate que non seulement les mots pertinents sont concentrée sur les extrémités de l'axe mais aussi qu'une grande partie des mots sont concentrés au centre. Ainsi nous pouvons supposer que favoriser les valeurs extrêmes de l'axe représente un filtre à la fois quantitatif et qualitatif.

Cette étude préliminaire valide, en tout cas pour ce qui est du corpus DUC 2004, la propriété selon laquelle plus la projection du plongement d'un mot en une dimension est éloignée de l'origine, plus il est probable que ce mot figure dans le résumé de référence. Nous présentons dans la Section 4 une méthode de résumé qui exploite cette propriété.

4 Méthode proposée

Nous présentons ici les différentes étapes de notre approche non supervisée permettant d'extraire les phrases les plus représentatives des documents. La Figure 2 représente schématiquement les grandes étapes de notre approche, qui vise à exploiter la propriété découverte et décrite en §3.

a) Afin de traiter des ensembles de documents de longueurs variables et potentiellement constitués de plusieurs milliers de mots, nous utilisons une représentation vectorielle statique par plongement de mots. Alors, un document est représenté par un ensemble de vecteurs denses. **b)** Nous appliquons une réduction de dimension extrême réduisant les plongements de mots à une seule et unique dimension. **c)** Ensuite, les mots partageant des contextes similaires sont regroupés. Les mots étant représentés sur une dimension unique, identifier les

groupes revient à segmenter la dimension lorsque la distance entre deux mots contigus est considérée comme trop grande. **d)** A partir des groupes de mots obtenus, nous définissons un score permettant d'attribuer un poids d'importance à chacun des groupes. **e)** Enfin, une fois les groupes de mots obtenus et pondérés, nous utilisons un *solveur* de contrainte permettant d'extraire les phrases maximisant les meilleurs mots selon un score défini tout en évitant que des mots pertinents mais trop proches (appartenant aux mêmes groupes) soient redondant dans l'ensemble des phrases extraites. Le *solveur* permet alors de maximiser à la fois la centralité mais aussi la diversité des mots.

4.1 Représentation et réduction des documents

Les plongements de mots permettent de représenter efficacement les documents. En raison de nos ressources calculatoires limitées, nous utilisons des plongements de mot statiques pré-entraînés. Cela permet d'obtenir directement un vecteur pour chaque mot du document. Afin de pouvoir les visualiser, nous utilisons la méthode UMAP (McInnes et al., 2020) permettant une réduction de dimension extrême non linéaire en favorisant la préservation des distances locales entre vecteurs par rapport à la distance globale.

4.2 Regroupement

La représentation des mots étant continue et univariée, cela permet d'utiliser une méthode linéaire de regroupement des mots proches sur la dimension. Concrètement, la distance moyenne entre chaque paire de points contigus est calculée et définit alors un seuil de regroupement. En dessous de ce seuil, deux points contigus sont considérés appartenir au même groupe. Au dessus de ce seuil, deux points contigus sont considérés appartenir à deux groupes différents. Cette approche permet de regrouper les mots proches sans introduire de paramètre supplémentaire et s'adapte quelque soit le nombre de mots présents dans le document.

4.3 Pondération

En fonction du document, la projection des mots sur une dimension peut être plus ou moins étendue. Afin d'homogénéiser le traitement des documents, la dimension est centrée et réduite. Empiriquement, l'importance d'un mot semblant être proportionnelle à sa distance au centre du « nuage » de mots, nous définissons une pondération ou distance à l'origine des mots d_O en ce sens :

$$d_O(x_m) = e^{(|x_m|)}$$

avec x_m la position du mot m sur la dimension réduite. Afin de pouvoir appliquer la pondération à un ensemble de mots du même groupe, celle-ci est alors définie comme la moyenne des pondérations des mots du groupe.

Des travaux proches comme (Gillick et Favre, 2009) définissent l'importance d'un mot comme étant directement dépendante de sa fréquence au sein du document. Dans le but d'intégrer cet aspect des mots à notre modèle, nous proposons de rajouter une dimension fréquentielle à la pondération des groupes de mots trouvés. Les groupes sont interprétés comme des *topics* constitués de mots partageant des contextes très similaires. Nous définissons la fréquence tf d'un groupe de mots par la fréquence de son mot le plus représenté :

$$tf(g) = \max_{m}^{m \in g} tf_g(m)$$

où le tf_g d'un mot m est défini par son nombre d'occurrences au sein du groupe.

4.4 Extraction des résumés

Nous nous inspirons de l'optimisation linéaire en nombres entiers pour l'extraction de résumés introduite avec ILP (Gillick et Favre, 2009). Celle-ci permet de maximiser la centralité des mots mais aussi la diversité de ceux-ci parmi les mots les plus fréquents. Nous utilisons l'optimisation linéaire mais en valeurs réelles [1] afin d'exploiter notre pondération des mots. Plus précisément, nous pondérons les mots par groupe. Aussi tout le vocabulaire du document est remplacé par les groupes d'appartenance des mots. Ainsi par exemple les mots m_1 et m_2 appartenant au groupe g_i sont remplacés tout deux par g_i dans les phrases les contenant et associés à leur pondération. L'optimisation consiste alors à retourner les phrases maximisant la pondération tout en minimisant leurs répétitions.

5 Evaluation

5.1 Données d'évaluation

Nous évaluons notre méthode sur le corpus DUC 2004. Il s'agit d'un corpus de résumé multidocument de *news* publié par le NIST à l'occasion de la campagne d'évaluation de *Document Understanding Conference*. Les documents sont constitués de plus de 6500 mots dans 260 phrases en moyenne. Chaque jeu de documents est associé à un résumé de référence rédigé par un humain. Le but est d'extraire des résumés de 9 phrases.

Nous utilisons les métriques ROUGE Lin (2004), fondées sur la comparaison de ngrammes entre résumés à évaluer et résumé(s) de référence. Nous présentons les résultats de ROUGE-1 (comparaison d'unigrammes), ROUGE-2 (comparaison de bigrammes) et ROUGE-L (plus longue sous-séquence commune).

5.2 Modèles de l'état de l'art testés

Nous comparons notre méthode à différents modèles de l'état de l'art :

Lead : consiste à extraire simplement les premières phrases du document, supposées porter une grande partie des informations pertinentes. **TextRank** : une méthode à base de graphe fondé sur les similarités entre mots (Mihalcea et Tarau, 2004). **SummPip** : une méthode à base de graphe intégrant différentes modalités, puis de partitionnement (Zhao et al., 2020). **SummVD** : une méthode fondée sur la réduction de dimension pour trouver des sujets latents vis-à-vis desquels les mots sont évalués (Shenouda et al., 2022). **ILP** : une méthode qui considère le résumé comme un problème d'optimisation et utilise la programmation linéaire en nombres entiers pour le résoudre (Gillick et Favre, 2009). **ILP-sem** : une variante de la méthode précédente, qui intègre une phase de *clustering* des phrases pour limiter la redondance dans le résumé (Mnasri et al., 2016).

1. https ://www.gnu.org/software/glpk/

	Métriques d'évaluation		
Méthode	ROUGE-1	ROUGE-2	ROUGE-L
Lead	30.66	08.36	14.73
TextRank	24.41	08.32	13.44
SummPip	36.30	08.47	-
SummVD	37.80	10.15	16.43
ILP	37.41	08.55	14.59
ILP-sem	38.28	10.17	15.15
REDIRE	**41.68**	**11.11**	**16.91**

TAB. 1 – Résultats obtenus sur tous les systèmes de l'état de l'art et l'approche proposée

	Métriques d'évaluation		
plongement de mot	ROUGE-1	ROUGE-2	ROUGE-L
FastText	**41.68**	**11.11**	**16.91**
Glove	41.11	10.49	16.67

TAB. 2 – Impact de différents plongements de mots

Toutes ces méthodes sont présentées en Section 2 et ont comme point commun leurs bonnes performances qui permettent de les placer, selon les évaluations des auteurs, au même niveau que les méthodes supervisées les plus récentes.

5.3 Résultats

Le tableau 1 présente les scores ROUGE de REDIRE et des différentes méthodes présentées en Section 5.2. On peut voir que REDIRE surclasse toutes les autres méthodes, quelle que soit la métrique, sur DUC 2004. Sur la métrique ROUGE-2, la plus corrélée des trois aux jugements humains selon Graham (2015), REDIRE dépasse la deuxième meilleure méthode, SummVD, d'un point ROUGE, soit environ 10% de bigrammes présents dans les résumés de référence extraits en plus.

5.4 Étude ablative

Dans cette Section nous évaluons l'influence des principaux paramètres du modèle. Le tableau 2 illustre l'impact des plongements de mots pré-entraînés sur la performance du modèle.

Le tableau 3 illustre l'impact des regroupements de mots sur la performance globale du modèle. L'écart important montre que les groupes trouvés sont cohérents et doivent permettre de limiter la redondance. Le tableau 4 illustre l'influence de la pondération des cluster sur la performance du système. On note que le meilleur résultat est la combinaison de la géométrie et de la fréquence.

avec et sans clustering	Métriques d'évaluation ROUGE-1	ROUGE-2	ROUGE-L
avec et score tf_{Max}	41.65	10.79	16.91
sans et score tf	37.41	08.55	14.59

TAB. 3 – Impact du regroupement de mots

score de cluster	Métriques d'évaluation ROUGE-1	ROUGE-2	ROUGE-L
$d_O \times tf_{Max}$	**41.68**	**11.11**	**16.91**
tf_{Max}	41.65	10.79	16.91
d_O	40.56	10.07	16.42

TAB. 4 – Impact des différents scores de clusters

6 Conclusions et perspectives

Nous avons présenté l'approche REDIRE permettant d'extraire des résumés à partir de documents longs. Grâce à une réduction extrême de dimension nous avons pu exploiter des propriétés géométriques des plongements de mots nous permettant d'obtenir de très bon résultats par rapport à un ensemble d'approches non supervisées. Nous avons montré empiriquement comment les mots spécifiques des documents présents dans les résumés pouvaient être retrouvés à partir de la projection des mots des documents. En perspective, nous voudrions étendre notre étude sur plusieurs corpus afin d'en évaluer plus largement l'efficacité. Il serait également intéressant de tester l'approche sur des plongements de mots dynamiques afin de tester la généralisation de l'approche.

Références

Gillick, D. et B. Favre (2009). A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, Boulder, Colorado, pp. 10–18. Association for Computational Linguistics.

Graham, Y. (2015). Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 128–137. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, Barcelona, Spain, pp. 74–81. Association for Computational Linguistics.

Manning, C. D., P. Raghavan, et H. Schütze (2008). *Introduction to Information Retrieval*. Cambridge, UK : Cambridge University Press.

McInnes, L., J. Healy, et J. Melville (2020). Umap : Uniform manifold approximation and projection for dimension reduction.

Mihalcea, R. et P. Tarau (2004). Textrank : Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, pp. 404–411. Association for Computational Linguistics.

Mikolov, T., E. Grave, P. Bojanowski, C. Puhrsch, et A. Joulin (2018). Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Mimno, D. et L. Thompson (2017). The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2873–2878. Association for Computational Linguistics.

Mnasri, M., G. de Chalendar, et O. Ferret (2016). Intégration de la similarité entre phrases comme critère pour le résumé multi-document (integrating sentence similarity as a constraint for multi-document summarization). In L. Danlos et T. Hamon (Eds.), *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. Volume 2 : TALN (Posters), Paris, France, July 4-8, 2016*, pp. 482–489. AFCP - ATALA.

Padmakumar, V. et H. He (2021). Unsupervised extractive summarization using pointwise mutual information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, Online, pp. 2505–2512. Association for Computational Linguistics.

See, A., P. J. Liu, et C. D. Manning (2017). Get to the point : Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Vancouver, Canada, pp. 1073–1083. Association for Computational Linguistics.

Shenouda, G., A. Bossard, O. Ayoub, et C. Rodrigues (2022). SummVD : An efficient approach for unsupervised topic-based text summarization. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, Online only, pp. 501–511. Association for Computational Linguistics.

Zhao, J., M. Liu, L. Gao, Y. Jin, L. Du, H. Zhao, H. Zhang, et G. Haffari (2020). Summpip : Unsupervised multi-document summarization with sentence graph compression. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, New York, NY, USA, pp. 1949–1952. Association for Computing Machinery.

Summary

This paper presents an unsupervised automatic summarization model capable of extracting the most important sentences from a corpus. To extract sentences in a summary, we use pre-entrained word embeddings to represent the documents. From this thick cloud of word vectors, we apply an extreme dimension reduction to identify important words, which we group by proximity. Sentences are extracted using linear optimization to maximize the information present in the summary. We evaluate the approach on large documents and present very encouraging initial results.