Détection de phrases comparatives pour le résumé de comparaison grâce à l'apprentissage profond.

Valentin Nyzam
LIFAT, 64 avenue Jean Portalis, 37200 Tours - France
valentin.nyzam@univ-tours.fr

RÉSUMÉ

Le résumé de comparaison a pour objectif de mettre en parallèle des informations comparatives et importantes de deux jeux de documents traitant d'une même thématique. Cela permet ainsi d'ajouter une nouvelle dimension de compréhension en analysant divers scénarios lors d'une même lecture. Dans le cadre de la génération d'un tel résumé, nous présentons une nouvelle méthode de détection de phrases comparatives utilisant l'apprentissage profond. En raison de l'absence de ressources disponibles pour cette tâche, nous avons composé un corpus d'évaluation. Sur ce corpus, notre méthode affiche les meilleurs résultats comparés à un ensemble de mesures de similarité sémantique classique.

ABSTRACT

Comparative sentence detection for comparative summarization using deep learning.

Comparative summarization seek to compare and contrast important information from two sets of documents dealing with the same topic. This task adds a new understanding's dimension by analyzing different scenarios at the same time. To generate such a summary, we present a new method for comparative sentence detection using deep learning. Due to the lack of available resources for this task, we built an evaluation corpus. On this corpus, our method shows the best results compared to a set of classical semantic similarity measures.

MOTS-CLÉS : détection de phrase comparative, apprentissage profond, similarité sémantique, résumé de comparaison.

KEYWORDS: comparative sentence detection, deep learning, semantic similarity, comparative summarization.

1 Introduction

Le résumé comparatif s'inscrit, de la même manière que pour le résumé d'actualités, dans le contexte actuel du nombre grandissant de données textuelles disponibles. Le résumé comparatif d'actualités cherche à comparer une paire d'ensembles de documents traitant de deux sujets différent au sein d'une même thématique – *ex : Tremblement de terre à Haïti / au Chili*. Pour cela, il est composé de deux "blocs", chacun résumant un seul groupe de documents. Les résumés comparatifs doivent alors être composés de phrases qui transmettent des informations à la fois représentatives de la thématique globale mais également comparatives entre les deux sujets des deux ensembles de documents.

En effet, un résumé d'actualité classique doit respecter trois caractéristiques principales :

- **Informativité** : Un résumé doit contenir un maximum de fragments textuels informatifs.
- **Redondance** : La redondance va influencer plus particulièrement les résumés multidocument. Il est en effet plus rare de rencontrer de la redondance au sein d'un seul et même document.
- **Longueur** : Un résumé est soumis à une contrainte de longueur.

Le résumé comparatif nécessite de respecter une propriété supplémentaire : la *comparativité*, qui exprime la comparaison. Une comparaison se compose de quatre éléments : le **comparé** (i.e. l'objet que l'on compare), le **comparant** (i.e. l'objet avec lequel est effectué la comparaison), l'**aspect** (i.e. l'échelle sur laquelle sont comparés le comparé et le comparant) et le **résultat** (i.e. le prédicat qui détermine les positions du comparé et du standard) (Huang *et al.*, 2011). Il est important de préciser que nous nous intéressons ici exclusivement à la comparaison factuelle et non à la comparaison rhétorique.

Une comparaison peut être explicitement exprimée dans une phrase comparative à l'aide de connecteurs linguistiques (comme, tel que, pareil que, plus que, ...). Elle peut également être présenter implicitement dans une section de texte qui décrit les caractéristiques individuelles de chaque objet point par point. Par exemple :

Haïti est un pays extrêmement pauvre.
Le Chili est un pays riche.

suggère ainsi que "le Chili est plus riche qu'Haïti". En effet, l'aspect comparatif est la richesse qui apparaît ici avec l'utilisation des adjectifs " pauvre" et "riche" appartenant à la même thématique. Le Chili se place ainsi haut sur l'échelle de comparaison de richesse et Haïti bas (riche $>$ pauvre). Le résultat est donc similaire à : *le Chili est supérieur à Haïti sur le thème de la richesse*.

Les méthodes de résumés comparatif vont ainsi chercher à extraire de telles paires de composants textuels (que nous identifierons dans la suite comme une "paire comparative"). La plus grande difficulté de génération d'un résumé de comparaison va ainsi être l'identification de telles paires comparatives. Dans cette étude, j'évalue ainsi un ensemble de méthode permettant une telle identification.

2 Travaux associés

Ils existent de nombreux travaux utilisant la comparaison mais celle-ci a été très peu étudié en tant que telle.

Dans le cadre du résumé comparatif d'opinions (ou résumé contrastif), Jindal & Liu (2006a,b) utilisent simplement la classe grammaticale des mots ainsi qu'un ensemble de règles afin d'identifier puis d'extraire des relations comparatives avec une F-mesure de 72%. Zhai *et al.* (2004); Wang *et al.* (2009) ont étudié la tâche de fouille de données comparatives. Ils utilisent des modèles probabilistes génératifs afin d'identifier les mots très différents dans un ensemble de documents. Afin d'aller plus loin dans l'utilisation des modèles probabilistes, les auteurs Campr & Ježek ont étudié l'utilisation de la méthode LSA (Campr & Ježek, 2012) et de la méthode LDA (Campr & Jezek, 2013) ainsi que des mesures de similarité simple (similarité cosinus et euclidienne) afin de générer des résumés comparatifs. Après avoir testé différents modèles probabilistes, les chercheurs ont étudié des modèles à base de graphe (Wan *et al.*, 2011) ainsi qu'une mesure de similarité sémantique (Huang *et al.*, 2011) : la similarité WordNet (Pedersen *et al.*, 2004). Plus récemment, des études utilisant l'apprentissage automatique ont commencé à apparaître malgré le peu de données d'apprentissage disponible. Duan & Jatowt (2019) étudient ainsi le résumé de comparaison temporel. Leur modèle cherche à apprendre une transformation orthogonale entre des collections d'articles de presses distants temporellement. Bista (2019) étudie le résumé de comparaison multimodal (textes, images et vidéos, ...). Il assimile alors ce problème à un problème de classification.

3 Méthodes

Dans premier temps, il est nécessaire de mettre en avant les diverses représentations des mots qui seront utilisées : une représentation Tf-Idf classique, les signatures thématiques ainsi qu'une représentation en espace vectorielle continu. Associé à ces représentations, cette étude évalue un ensemble de 6 mesures de similarité : la similarité cosinus, la distance euclidienne, la similarité de Jensen-Shannon, la similarité WordNet (Pedersen *et al.*, 2004), la métrique de Wasserstein (Kusner *et al.*, 2015) ainsi qu'une mesure fondée sur l'apprentissage automatique.

Du fait du manque de données disponibles, il est difficile d'utiliser des méthodes d'apprentissage supervisées. Néanmoins, lors de notre travail, nous avons empiriquement déterminé que la tâche de détection de paires comparatives peut être rapprochée de celle de détection de paraphrase. En effet, de la même manière que les phrases comparatives, les paraphrases partagent une thématique identique. La mesure fondée sur l'apprentissage automatique est ainsi fondée sur la méthode DIIN (Gong *et al.*, 2017) initialement proposée pour la détection de paraphrase. Ce réseau de neurone est composé d'une suite de couche ayant chacune un objectif particulier : la couche d'encodage, – qui encode les représentations vectorilles des mots en incorporant de l'information contextuelle – la couche d'interaction, – qui crée le tenseur d'interaction entre les mot en utilisant les caractéristiques des phrases d'entrées – la couche d'extraction des caractéristiques, – qui a pour objectif d'extraire les caractéristiques sémantiques détectable dans le tenseur d'interaction précédent – une couche de sortie, – qui décode les caractéristiques extraites afin de calculer une prédiction. Ainsi, dans le cadre de l'inférence linguistique, la couche de sortie prédit la confiance sur chaque classe. Gong *et al.* utilisent pour cela une simple couche linéaire.

Cette méthode est alors entraînée sur le corpus d'inférence et de paraphrase (Wang *et al.*, 2018) puis finalement affinée sur le corpus de paraphrase STS-B (Agirre *et al.*, 2012) adaptée pour la détection de paires comparatives. Nous ajoutons également la possibilité d'intégrer des indices de surfaces supplémentaires à cette méthode - analyse grammaticale et détection d'entité nommées.

4 Corpus

À ma connaissance, il n'existe pas de corpus permettant d'évaluer la comparaison. L'objectif final de cette étude étant d'extraire des paires de phrases comparatives, il a été nécessaire de construire manuellement un corpus de paire de phrases. Chaque paire de phrases est alors annoté comme comparative ou non. Ce corpus est composé d'environ 3000 paires de phrases issues d'articles de journaux.

5 Évaluation

Les poids Tf-Idf sont initialement calculés sur le corpus Reuters disponible dans le package python nltk. Les distributions des méthodes LDA sont obtenues sur ce même corpus. Nous utilisons de plus des représentations en espace vectoriel continu pré-entraîné : Fasttext [1]. Nous utilisons cette représentation car les auteurs [2] mettent à disposition un ensemble de modèles multilingues. Cela peut ainsi permettre une transition plus simple des méthodes utilisant cette représentation vers le multilinguisme. Les résultats présentés dans le tableau 1 sont alors obtenus.

1. https ://github.com/facebookresearch/fastText
2. e.g. Facebook

Modélisation des phrases	Mesure de similarité	Précision	Rappel	F-Mesure
Tf-Idf	Similarité Cosinus	68.7%	88.1%	77.2%
Tf-Idf	Distance Euclidienne	69.3%	88.2%	77.6%
Tf-Idf	Similarité WordNet	77.7%	91.2%	83.9%
LDA	Divergence Jensen Shannon	72.2%	91.2%	80.6%
WE moyen	Similarité Cosinus	76.3%	90.5%	82.8%
WE moyen	Distance Euclidienne	76.9%	90.5%	83.2%
WE	Similarité WMD	79.6%	92.6%	85.6%
WE	DIIN	**82.1**%	92.8%	86.4%
WE + indices de surface	DIIN	**82.1**%	**93.7**%	**87.5**%

TABLEAU 1 – Résultats obtenus pour notre étude sur la détection de la comparativité.

Nous pouvons ainsi remarquer de manière globale que, comme pour la tâche de détection de paraphrase, le rappel est bien plus important que la précision. Les différentes méthodes de détection de la comparativité ont ainsi tendance à plus facilement détecter les paires comparatives que les paires non-comparatives. En effet, une mesure de rappel élevé signifie que toutes les paires comparatives sont identifiées (i.e. il y a peu de faux négatifs, de paires comparatives identifiées comme non-comparatives). Tandis qu'une mesure de précision plus faible signifie que des paires non-comparatives sont plus souvent identifiés comme comparative.

Ainsi, la représentation Tf-Idf obtient les moins bons résultats. En effet, cette représentation ne transporte que des informations statistiques sur la fréquence des termes dans un corpus et aucune sémantique. Il est néanmoins intéressant de remarquer que l'utilisation de la distance euclidienne améliore très légèrement les résultats.

La similarité WordNet améliore grandement ces résultats. En effet, cette similarité est basée sur les relations entre les concepts WordNet (par exemple, bus et voiture seront proches contrairement à bus et chien) ce qui est une sémantique relativement précise. L'un des désavantage de cette méthodes vient de sa dépendance au thésaurus WordNet. Il n'est ainsi pas simple de généraliser cette méthode au multilinguisme sans utiliser la traduction automatique.

L'utilisation des signatures thématiques avec la méthode LDA améliore les résultats obtenus avec Tf-Idf et la distance euclidienne mais reste en deçà des résultats obtenus par la similarité WordNet, de même pour les méthodes utilisant les représentations vectorielles moyennes (*WE moyen*). Ces résultats moyens peuvent s'expliquer par le fait qu'initialement, ce sont des représentations pour les mots. Celles-ci sont alors moyennées sur l'ensemble de la phrase, ce qui entraîne alors une perte d'information. En effet, la sémantique associée au vecteur moyen n'est pas forcément représentative de la combinaison des vecteurs des mots.

La similarité WMD remédie à ce problème en proposant de représenter la phrase comme un sac de mots, chacun associé à son vecteur (Kusner *et al.*, 2015). Les résultats s'améliorent bien dans ce cas là, ce qui valide le fait que représenter la phrase comme la moyenne des vecteurs des mots qui la compose n'est pas une bonne approche.

En dernier lieu, la méthode neuronale DIIN obtient de très bons résultats. En effet, sa couche d'extraction des caractéristiques permet au réseau de neurones de découper différentes portions de l'espace vectoriel d'entrée et de focaliser son attention sur les parties les plus intéressantes de la phrase. L'apport d'indices de surface, classe grammaticale et détection d'entités nommées, permet de plus d'améliorer simplement les résultats de cette méthode. Il serait ainsi intéressant d'ajouter ces indices de surface comme caractéristiques d'entrée à une méthode classique de similarité afin de vérifier si l'impact serait aussi important. Nous n'avons malheureusement pas encore testé une telle variante. Nous pouvons finalement remarquer que le pré-apprentissage du réseau de neurones de la méthode DIIN sur un corpus dédié à l'inférence a finalement bien fonctionné.

Références

AGIRRE E., CER D., DIAB M. & GONZALEZ-AGIRRE A. (2012). Semeval-2012 task 6 : A pilot on semantic textual similarity. In ** SEM 2012 : The First Joint Conference on Lexical and Computational Semantics–Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, p. 385–393.

BISTA U. (2019). Comparative summarisation of rich media collections. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, p. 812–813.

CAMPR M. & JEŽEK K. (2012). Comparative summarization via latent semantic analysis. *Department of Computer Science and Engineering University of West Bohemia*.

CAMPR M. & JEZEK K. (2013). Comparative summarization via latent dirichlet allocation. In *Dateso*, p. 80–86 : Citeseer.

DUAN Y. & JATOWT A. (2019). Across-time comparative summarization of news articles. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, p. 735–743.

GONG Y., LUO H. & ZHANG J. (2017). Natural language inference over interaction space. *arXiv preprint arXiv :1709.04348*.

HUANG X., WAN X. & XIAO J. (2011). Comparative news summarization using linear programming. In *Proceedings of the 49th Annual Meeting of the ACL*, p. 648–653.

JINDAL N. & LIU B. (2006a). Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 244–251 : ACM.

JINDAL N. & LIU B. (2006b). Mining comparative sentences and relations. In *AAAI*, volume 22, p. 1331–1336.

KUSNER M., SUN Y., KOLKIN N. & WEINBERGER K. (2015). From word embeddings to document distances. In *International Conference on Machine Learning*, p. 957–966.

PEDERSEN T., PATWARDHAN S. & MICHELIZZI J. (2004). Wordnet : : Similarity : measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, p. 38–41 : Association for Computational Linguistics.

WAN X., JIA H., HUANG S. & XIAO J. (2011). Summarizing the differences in multilingual news. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, p. 735–744 : ACM.

WANG A., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. (2018). GLUE : A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 353–355, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5446.

WANG D., ZHU S., LI T. & GONG Y. (2009). Comparative document summarization via discriminative sentence selection. In *Proceedings of the 18th ACM conference on Information and knowledge management*, p. 1963–1966.

ZHAI C., VELIVELLI A. & YU B. (2004). A cross-collection mixture model for comparative text mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 743–748 : ACM.