Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Étude sur le résumé comparatif grâce aux plongements de mots

Valentin Nyzam Aurélien Bossard
Laboratoire d'Informatique Avancée de Saint-Denis, Université Paris 8
2 rue de la Liberté, 93526 Saint-Denis
valentin.nyzam@iut.univ-paris8.fr, aurelien.bossard@iut.univ-paris8.fr

RÉSUMÉ

Dans cet article, nous présentons une nouvelle méthode de résumé automatique comparatif. Ce type de résumé a pour objectif de permettre de saisir rapidement les différences d'information entre deux jeux de documents. En raison de l'absence de ressources disponibles pour cette tâche, nous avons composé un corpus d'évaluation. Nous présentons à la fois la méthodologie de son élaboration ainsi que le corpus lui-même. Notre méthode utilise les avancées récentes dans le calcul de similarité entre phrases afin de détecter les informations comparatives. Nous montrons que sur ce corpus, notre méthode est comparable en termes de qualité de résultats à une méthode de l'état de l'art, tout en réduisant d'un facteur dix le temps de calcul, la rendant donc exploitable dans le cadre de l'aide à l'analyse de documents.

ABSTRACT

Comparative summarization study using word embeddings.

This paper introduces a new method for comparative automatic summarization. This kind of summary allows to quickly identify the differences between two sets of documents. Because of the lack of evaluation resources, we built an evaluation corpus. We present both the methodology used to build this corpus, and the corpus itself. Our method makes use of recent advances in sentence similarity computation in order to detect comparative information. We show that on that corpus, our method compares to a state-of-the-art method in terms of quality and divides the computation time by a factor ten. This increase in computation speed makes automatic comparative summarization usable for support for document analysis.

MOTS-CLÉS : Résumé automatique, résumé comparatif, plongements de mots.

KEYWORDS: Automatic summarization, comparative summarization, word embeddings.

1 Introduction

Étant donné le nombre d'actualités disponibles en ligne, l'analyse automatique des articles de presse est devenue un enjeu important. Dans ce cadre, nous nous intéressons au résumé comparatif qui permet de regrouper les informations connexes au sein d'un groupe de documents qui traitent d'un même type de sujet pour les présenter à l'utilisateur sous la forme de courts résumés. Une telle présentation analytique est une réelle plus-value pour le lecteur du résumé. Elle met en effet en avant les informations comparables au sein de plusieurs documents, ce qui peut s'apparenter au résumé guidé par les aspects, mais de manière non supervisée. Cependant, cette tâche représente un défi pour le traitement automatique de la langue.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 95–107.
HAL : hal-02784758.

Il est en effet nécessaire d'avoir une connaissance approfondie des sujets traités par les documents afin de trouver de bonnes comparaisons. C'est pourquoi l'analyse manuelle prend généralement beaucoup de temps et de travail. La génération automatique d'un résumé qui mette automatiquement en évidence les informations connexes entre les sujets d'actualité serait ainsi une aide primordiale afin de pouvoir effectuer des analyses de manière plus simple et efficace. Dans la suite de cet article, nous appelons "informations comparatives" deux informations qui peuvent être rapprochées l'une de l'autre d'après le thème qu'elles portent.

Le résumé comparatif d'actualités cherche ainsi à comparer une paire de documents ou une paire d'ensembles de documents. Le résumé généré est donc composé de deux "blocs", chacun résumant un seul document (ou un seul groupe de documents). De plus, les résumés doivent être composés de phrases qui transmettent des informations à la fois comparatives et représentatives de chaque document ou ensemble de documents.

Nous présentons ici une nouvelle méthode de résumé comparatif fondée sur les récentes avancées en sémantique computationnelle : les plongements de mots et la *Word Mover Distance* (WMD) (Kusner *et al.*, 2015) afin de détecter les phrases sémantiquement proches pour la comparaison. Les résultats expérimentaux démontrent l'efficacité de ce modèle en termes de temps d'exécution ainsi que de qualité d'identification et de résumé des comparaisons.

Dans un premier temps, nous présentons les travaux associés à la tâche du résumé comparatif. Nous expliquons ensuite notre méthode de résumé comparatif, ainsi que les méthodes d'évaluation de la comparativité et de l'informativité. Nous présentons alors le contexte de notre expérience ainsi que les résultats obtenus, puis les conclusions et suites que nous envisageons à cette étude.

2 Définition

Wang *et al.* (2009) sont les premiers à avoir défini le résumé comparatif. Ils définissent le résumé comparatif comme suit : *"Étant donnée une collection de groupes de documents, le résumé de comparaison consiste à générer un court résumé composé des différences entre ces documents en extrayant les phrases les plus discriminantes dans chaque groupe de documents."*. En n'extrayant que des phrases discriminantes, le résumé comparatif s'écarte de la définition d'une comparaison car il y manquera sûrement les aspects comparatifs. Il se rapproche en revanche de la tâche traditionnelle de résumé automatique, si ce n'est qu'au lieu d'extraire les informations les plus discriminantes d'un document ou groupe de documents, on extrait les informations les plus discriminantes d'un groupe de documents vis-à-vis des autres. En effet, l'extraction des seules phrases discriminantes sans tenir compte d'une thématique commune implique que le résumé comparatif perd alors sa cohérence et son sens. Wang *et al.* (2009) fournissent un exemple de phrases discriminantes en utilisant leur méthode *Discriminative Sentence Selection*. Le tableau 1 présente un résumé automatique tiré de l'article qui présente ce travail. Bien que l'on ne dispose pas des articles d'origine, on peut constater que cet exemple constitue difficilement un résumé. Il est même compliqué de comprendre quelle en est l'information principale. De plus, les phrases ainsi extraites de leur contexte sont difficilement interprétables, et l'intérêt de cette méthode vis-à-vis d'un utilisateur final est donc discutable.

Huang *et al.* (2014) étendent la définition précédente en déclarant que : *"Un résumé comparatif doit contenir une combinaison de deux composants (provenant de différents ensembles de documents), chacun lié au même sujet."*. Huang *et al.* (2014) fournissent un exemple de résumé de référence (écrit

ID	Discriminative sentences
1	There is no cold war, there is no Saddam. Lebanon has also changed.
2	If hiring rap sheet-free intelligent people means they won't hire a black applicant for another five years.
3	He should drop the case against the lacrosse players but not the sexual assault case itself.
4	Rahman, who is about 41 years old, converted from islam to christianity over 16 years ago.
5	To be totally honest with you, we believed that there may have been a classified annex.
6	I suspect that his position reflects conventional wisdom among the Chinese military establishment.
7	In both the short and long term what those displaced by hurricane Katrina need most is money.

TABLE 1 – Extraction de phrases discriminantes par Wang *et al.* (2009).

par un humain) présenté dans le tableau 2, qui compare la coupe du monde de football 2006 à celle de 2010. Les phrases qui traitent du même thème ne sont pas mises en vis-à-vis. Toutefois, les thèmes sont ordonnés de la même manière au sein des deux résumés, et on peut dès lors identifier les thèmes principaux mis en avant par le rédacteur. Le rapprochement entre ce type de résumé comparatif et le résumé guidé par les aspects est évident. La lecture de ces résumés apporte des éléments concrets de compréhension et fait bien ressortir les spécificités des jeux de documents source.

Nous voyons donc le résumé comparatif de la même manière que Huang *et al.* (2014), à savoir la mise en parallèle de phrases issues de deux ensembles différents afin de faire apparaître les similarités thématiques et les différences des informations à l'intérieur de chaque thème qui font la spécificité de chaque jeu de documents.

Ensemble de documents A	Ensemble de documents B
Italy claimed a fourth world title in a penalty shoot-out victory over France after the two sides finished a goal apiece following extra-time in Berlin's Olympic Stadium on Sunday	Spain have won the 2010 FIFA World Cup South Africa final, defeating Netherlands 1-0 with a wonderful goal from Andres Iniesta deep into extra-time.
France captain Zinedine Zidane won the Golden Ball award for the tournament's best player	Uruguay star striker Diego Forlan won the Golden Ball Award as he was named the best player of the tournament at the FIFA World Cup 2010 in South Africa
Lukas Podolski was named the inaugural Gillette Best Young Player by FIFA's TSG after scoring three goals and contributing boundless energy to Germany's enthralling FIFA World Cup campaign	German youngster Thomas Mueller got double delight after his side finished third in the tournament as he was named Young Player of the World Cup by the FIFA Technical Study Group (TSG) and he also won the Golden Boot Award for the tournament's top-scorer.
Germany striker Miroslav Klose was the Golden Shoe winner for the tournament's leading scorer	
Germany's minister of economics and technology, Michael Glos, says he is confident the World Cup will boost the economy.	The net economic benefit from hosting the World Cup for South Africa, in terms of current and future tourism impact, is unclear
An average of 52,500 fans packed into the 12 stadiums for the 64 matches	South Africa will have five brand new state of the art football stadiums that seat an average of 50,400 spectators and five newly renovated stadiums that seat an average of 53,300
In Berlin, for example, police estimated that up to one million fans converged on the official Fan Fest public viewing venue in front of the Brandenburger Tor on Saturday to watch the host nation beat Sweden for a quarterfinal berth	In Berlin, about 3,50,000 people watched Germany at the FIFA fan fest on Wednesday night, while 56,836 people attended the fan fest in Durban
Television audiences for the 2006 FIFA World cupTM in Germany are being collated as the tournament progresses and it already looks as if they are heading for the record books	A global TV audience of more than 700 million watched Sunday's World Cup final, according to the tournament's organizers

TABLE 2 – Exemple d'un résumé de référence issu de Huang *et al.* (2014).

3 Travaux associés

Le résumé comparatif est étudié depuis de nombreuses années en linguistique et de nombreux travaux ont étudié la connotation, l'extension, les formes et les usages des comparaisons (Kennedy, 2007; Lerner & Pinkal, 2003). L'analyse comparative a été appliquée dans de nombreux domaines, et plusieurs sujets académiques connexes ont émergé, tels que la littérature linguistique (Anttila, 1972), la littérature (Prawer, 1973), l'histoire et la politique comparées. L'analyse comparative est également largement utilisée dans les applications web. De nombreux systèmes de commerce électronique fournissent des comparaisons de produits de base sur les prix et les fonctionnalités en se basant sur les données structurelles sous-jacentes.

Plus récemment, l'extraction d'informations comparatives à partir de données non structurées a attiré beaucoup d'attention. Plusieurs chercheurs proposent d'étudier la comparaison d'entités en utilisant des modèles linguistiques (Bao *et al.*, 2008) ou des mesures de similarités entre distributions de probabilités (Jain & Pantel, 2011; Liu *et al.*, 2007). Certains travaux tentent d'identifier des comparaisons linguistiques explicites et d'en extraire les éléments de comparaison (Jindal & Liu, 2006), d'autres cherchent à extraire des caractéristiques individuelles dans les phrases et à les faire correspondre (Kim & Zhai, 2009; Paul *et al.*, 2010; Zhai *et al.*, 2004).

D'autres encore utilisent des méthodes de résumé automatique conventionnelles telles que LSA, LDA, méthodes à base de graphe, ILP... L'intérêt des méthodes à base de LSA et LDA (Campr & Ježek, 2013) est la représentation des documents sous la forme de distributions de probabilités sur des thématiques. Les méthodes à base de graphes (Wan *et al.*, 2011) et de programmation linéaire en nombres entiers (Huang *et al.*, 2014) permettent de conserver simplement l'informativité présente dans les ensembles de documents, comme dans le cadre du résumé automatique classique.

Si la plupart de ces études se concentrent sur la comparaison des aspects communs au sein de phrases, il existe également certaines recherches axées sur la détection des informations uniques des sujets qui composent les documents (Wang *et al.*, 2012) ou de la nouveauté de ces informations.

La méthode utilisée par (Huang *et al.*, 2014) étant la plus proche de nos travaux, nous l'utiliserons comme baseline de comparaison dans nos expériences. Afin de générer les résumés, les auteurs utilisent à la fois la comparabilité, la centralité et la non-redondance dans une fonction objectif dont ils cherchent le maximum grâce à la programmation linéaire en nombres entiers (ILP). Leur méthode fait appel à une ressource sémantique exogène : WordNet (Pedersen *et al.*, 2004) afin de capturer les similarités entre des paires de concepts (unigrammes ou bigrammes) et ainsi déterminer le poids "comparatif" d'un ensemble de paires de concepts. Une telle méthode possède deux inconvénients majeurs :

— son coût de calcul : elle nécessite le calcul de similarité entre concepts dans un arbre et ce pour tous les concepts identifiés dans le corpus à résumer ;
— sa portabilité vers d'autres langues, bien que des ressources telles que BabelNet (Navigli & Ponzetto, 2012) répondent en partie à ce problème.

4 Méthode proposée

Dans notre étude, nous nous intéressons au résumé comparatif tel que défini dans (Huang *et al.*, 2011, 2014). Les auteurs tentent d'identifier les phrases comparables entre deux ensembles documents, i.e.

qui partagent un thème commun, tout en véhiculant une information différente. Il faut donc extraire des paires de phrases comparables depuis les deux ensembles de documents mais également identifier parmi ces paires comparables quelles sont celles qui véhiculent un thème important. Il faut en effet éviter de présenter au lecteur du résumé des informations non essentielles. La difficulté inhérente à la tâche consiste donc à trouver un compromis entre la comparabilité des phrases à extraire dans le résumé et l'aspect central des informatios qu'elles contiennent.

Nous construisons un résumé comparatif en extrayant des phrases sémantiquement similaires. Afin d'identifier les phrases similaires, nous utilisons une mesure de distance sémantique destinée aux plongements de mots : la *Word Mover Distance (WMD)* (Kusner *et al.*, 2015). La distance WMD mesure la dissimilarité entre deux documents (dans notre cas, des phrases) comme la distance minimale que les plongements de mots d'un document doivent "parcourir" pour atteindre les plongements de mots d'un autre document. Le calcul de cette distance consiste donc en une résolution d'un problème de transport.

Nous déterminons la comparabilité au niveau de la phrase afin d'améliorer le temps de traitement, contrairement à (Huang *et al.*, 2014) qui travaillent au niveau des concepts. En effet, utiliser les concepts est problématique : Les jeux de documents sont constitués d'entre trois et cinq milles concepts différents. Pour chaque résumé, il serait donc nécessaire de calculer en moyenne 15 millions de similarités sémantiques entre paires de concepts. En travaillant au niveau des phrases, nous limitons le nombre de paires de phrases à 100 000, ce qui semble plus réaliste d'un point de vue calculatoire.

Il nous faut également une mesure de centralité des phrases, c'est-à-dire de l'importance de l'information véhiculée par les phrases vis-à-vis du jeu de documents auquel elles appartiennent. Pour cette première approche du problème, et bien que d'autres mesures plus performantes existent (Gambhir & Gupta, 2017) nous avons choisi d'utiliser, la somme du tf.idf des concepts d'une phrase afin d'évaluer sa centralité.

Ces deux scores de comparativité (entre des paires de phrases) et de centralité (pour une phrase) sont normalisés puis combinés. Cette combinaison des deux scores constitue alors une fonction objectif destinée à représenter le compromis entre comparativité et centralité, et qui peut être intégrée à un algorithme d'optimisation afin de déterminer le résumé qui la maximise.

La fonction objectif utilisée est la suivante :

$$obj(sum) = (1 - \lambda)Score_{info}(sum) + \lambda Score_{comp}(sum) \tag{1}$$

avec $Score_{info}$ la fonction représentant l'informativité du résumé et $Score_{comp}$ la fonction représentant la comparativité du résumé.

Le score d'informativité du résumé est calculé simplement à l'aide d'une somme sur les TF-IDF des concepts présents :

$$Score_{info}(R) = \sum_{i=1}^{2} \sum_{k=1}^{|C_i|} w_{ij} \cdot présent(c_{ij}, R) \tag{2}$$

avec $C_i = \{c_{ij}\}$ l'ensemble des concepts présents dans l'ensemble de documents D_i. Un concept est défini comme un unigramme ou un bigramme. $présent(c_{ij}, R)$ représente la présence ou non du concept c_{ij} dans le résumé R. Chaque concept c_{ij} a un poids $w_{ij} \in \mathbb{R}$ calculé comme le score TF-IDF du concept sur l'ensemble de documents D_i :

$$w_{ij} = TF(c_{ij}, D_i) \cdot IDF(c_{ij}) \tag{3}$$

Le score de comparativité est calculé comme la somme des poids comparatifs présents dans le résumé. Un poids comparatif est le poids associé à la comparaison entre deux phrases issues des deux différents jeux de documents. Il se calcule comme la somme des similarités normalisées entre les phrases appartenant au résumé A et les phrases appartenant au résumé B, calculées avec la distance WMD.

$$Score_{comp}(R) = \sum_{s_{1j} \in R} \sum_{s_{2k} \in R} \frac{sim_{WMD}(s_{1j}, s_{2k})}{\max_{\forall m,n}(sim_{WMD}(s_{1m}, s_2 n))} \tag{4}$$

avec s_{ij} la phrase j de l'ensemble de documents i, R le résumé (composé du résumé de chacun des deux jeux de documents) et sim_{WMD} la similarité WMD calculée à partir de la *Word Mover Distance*. Une paire de phrases $< s_{1j}, s_{2k} >$ a ainsi un poids considéré comme égal à la distance WMD normalisée qui les sépare. Ce poids indique si la paire de phrases est porteuse d'une comparaison importante.

Nous appliquons en dernier lieu un algorithme d'extraction destiné à la résolution du problème du sac à dos (Pisinger *et al.*, 2007), qui a déjà été utilisé avec succès pour le résumé automatique (McDonald, 2007) avec la fonction objectif définie en équation 1.

5 Expérience

Dans cette section, nous décrivons l'expérience mise en place afin d'évaluer notre méthode. Nous y présentons notre corpus, la chaîne de traitement de notre méthode à des fins de reproductibilité, la *baseline* utilisée ainsi que la mesure utilisée pour l'évaluation. En dernier lieu, nous présentons les paramètres expérimentaux et les résultats de l'expérience.

5.1 Corpus

En raison de la nouveauté de la tâche du résumé comparatif, il n'existe pas de jeu de données disponible pour son évaluation. Huang *et al.* (2014); Campr & Ježek (2013) ont chacun créé un jeu de données de dix corpus, chaque corpus étant composé d'une paire d'ensembles de dix documents sur des sujets comparables. Sur le même principe illustré en figure 1, nous avons donc créé notre propre ensemble de corpus en anglais.

Nous avons d'abord choisi dix paires de sujets comparables, présentées en figure 3. Puis nous avons récupéré les dix premiers articles de presse liés à chaque sujet en utilisant un moteur de recherche populaire. Enfin, nous avons créé manuellement le résumé humain pour chaque paire de sujets. Le résumé humain a été créé en fonction d'instructions souples : "les deux résumés doivent contenir uniquement les informations comparables issues des deux jeux de documents, et sont chacun limités à 100 mots. Informations comparables : informations participant à un même thème à l'intérieur des jeux de documents.". Il est important de noter que chaque résumé de référence contient également deux "blocs" limités à 100 mots, chacun d'entre eux comportant les phrases de l'ensemble de documents A ou B. Un résumé se compose donc d'un maximum de 200 mots.

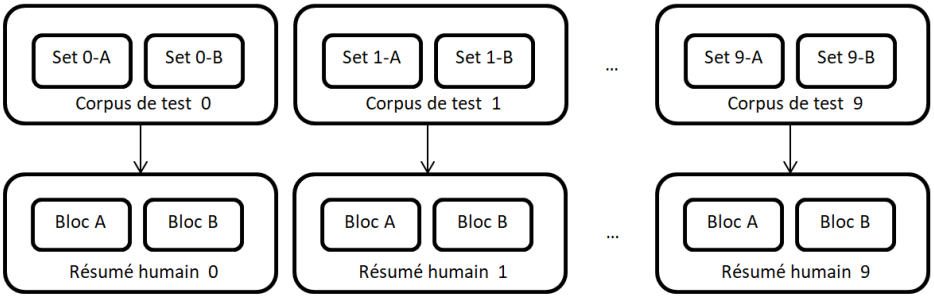

FIGURE 1 – Schéma illustrant la création du corpus.

5.2 Chaîne de traitement

La chaîne de traitement de notre méthode pour cette expérience est présentée dans la figure 2. Elle consiste tout d'abord dans les traitements classiques : tokenisation et segmentation en phrases, puis en une mise à jour, sur le corpus de résumé, des plongements de mots calculés au préalable sur un corpus issu de Wikipedia. Cette mise à jour spécifique au corpus d'évaluation vise à ajouter aux plongements de mots préexistants le vocabulaire spécifique qui n'y serait pas présent. Avoir une liste exhaustive des plongements de mots est en effet un prérequis nécessaire pour que la *Word Mover Distance* puisse délivrer des résultats satisfaisants.

Ces pré-traitements sont suivis des différents traitements visant à mettre en place la fonction objectif, suite à quoi l'algorithme d'optimisation de type résolution du problème du sac à dos décrit dans (Pisinger *et al.*, 2007) est lancé. Notre méthode, décrite en §4, est nommée SenWE-KP.

ID	Thématique	Sujets
1	Catastrophe naturelle	Tremblement de terre Haïti / Chili
2	Terrorisme	Attaque Paris / Nice
3	Politique	Élection présidentielle US / France
4	Politique	Mariage homosexuel US / France
5	Politique	Candidature Paris JO 2024 / 2012
6	Catastrophe naturelle	Feu Forêt Portugal / USA
7	Catastrophe naturelle	Inondation France / USA
8	Catastrophe naturelle	Tsunami Fukushima / 2004 Océan Indien
9	Catastrophe naturelle	Ouragan Irma / Katrina
10	Scandale	Affaire DSK / Affaire Weinstein

TABLE 3 – Paires de sujets comparables et leur thématique constituant l'ensemble de corpus d'évaluation.

5.3 Baseline

Nous utilisons notre implémentation de la méthode de Huang *et al.* (2014) comme méthode de référence (WordNet-ILP). Malheureusement, les auteurs n'ont pas fourni le code correspondant. Nous espérons notre implémentation la plus fidèle possible à leur méthode décrite en §3 ; elle associe une similarité sémantique entre concepts issue de WordNet avec une mesure de centralité pour obtenir une

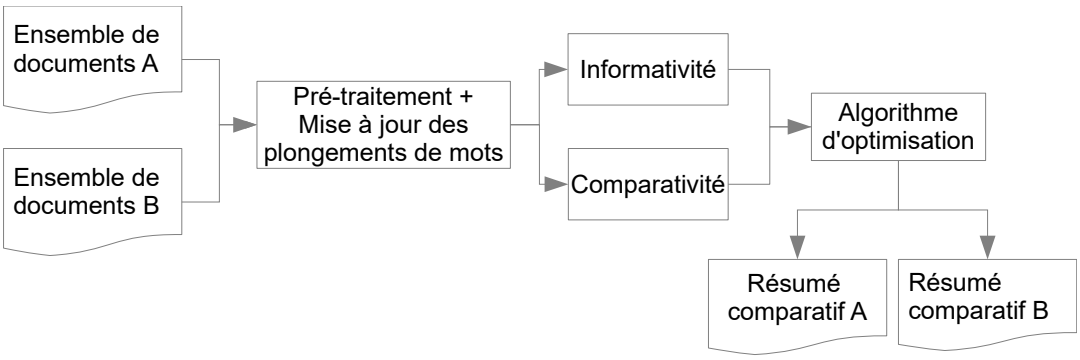

FIGURE 2 – Chaîne de traitement pour le résumé comparatif.

fonction objectif. Cette fonction objectif est maximisée grâce à un solveur de programmation linéaire en nombres entiers, guidé par des contraintes similaires à celles décrites dans (Gillick & Favre, 2009) qui interdisent toute redondance de concepts au sein du résumé généré.

5.4 Évaluation

L'évaluation d'un résumé reste toujours problématique. On peut catégoriser les méthodes d'évaluation en quatre catégories :

— automatiques sans référence (Louis & Nenkova, 2013) ;
— automatiques avec référence (ROUGE, BLEU a aussi été utilisé dans certaines études) (Lin, 2004) ;
— manuelles avec score automatique : Pyramid (Nenkova & Passonneau, 2004) ;
— manuelles entièrement subjectives.

Huang *et al.* (2014) ont proposé une méthode d'évaluation semblable à Pyramid mais adaptée au résumé comparatif. Il s'agit d'évaluer si les résumés automatiques traitent ou non des thèmes relevés dans les résumés manuels. Cependant, les évaluations manuelles étant coûteuses en temps humain, nous souhaitons les réserver pour les étapes futures de notre étude. Les évaluations automatiques sans références, quant à elles, sont inexploitables dans notre contexte. En effet, elles s'appuient sur la similarité entre le jeu de documents à résumer et les résumés à évaluer, par le biais des distributions de probabilité. L'objectif du résumé comparatif s'éloigne de celui du résumé "traditionnel" pour lequel la centralité, bien reflétée par les mesures de divergence entre distributions de probabilité, est la notion la plus importante. Nous avons donc opté pour les mesures ROUGE.

Nous évaluons les deux résumés créés A et B pour chaque ensemble de documents vis-à-vis respectivement des résumés humains A et B. Dans cette expérience, nous utilisons les paramètres ROUGE possédant la meilleur corrélation avec les scores humains d'après Graham (2015) :

— ROUGE-2 : Précision ROUGE-2 moyenne avec racinisation et suppression des mots vides [1]
— ROUGE-3 : F-Mesure ROUGE-3 moyenne sans racinisation et sans suppression des mots vides [2].

1. Arguments ROUGE-2 : *-n 2 -x -m -s -c 95 -r 1000 -f A -p 1 -t 0*
2. Arguments ROUGE-3 : *-n 3 -x -c 95 -r 1000 -f A -p 0.5 -t 0*

Résumé A (Chili) : Reuters now reports that 47 people have died in the quake. The earthquake struck at 3 :34 a.m. in central Chile, centered roughly 200 miles southwest of Santiago at a depth of 22 miles, the United States Geological Survey reported. More than 300 people were killed, according to Chile's Office of Emergency Management, and 15 are missing. A tsunami watch has been issued for Ecuador, Colombia, Panama, Costa Rica and Antarctica. No damage was expected from possibly stronger waves to follow, Ryan said. Chilean President Michele Bachelet said that altogether two million people had been affected.
Résumé B (Haïti) : The headquarters of the U. N. peacekeeping mission in Port-au-Prince collapsed, a U. N. official told CNN. "We stand ready to assist the people of Haiti," Mr. Obama said. Some said that they had been able to get through immediately after the earthquake. The World Bank forgave the country's $36 million balance in May. "I saw people under the rubble, and people killed. A "large number" of UN personnel were reported missing by the organisation. The quake was felt in the Dominican Republic, sending people in the capital Santo Domingo running on to the streets in panic.

FIGURE 3 – Résumé comparatif issu de la méthode WordNet-KP pour le sujet "Tremblement de terre Chili / Haiti 2010".

5.5 Paramètres expérimentaux

Tous les systèmes génèrent un résumé comparatif composé de deux résumés de 100 mots chacun. Toutes les expériences sont effectuées sur un processeur Intel Xeon à 2,20 GHz composé de 40 coeurs. Les résultats de l'évaluation sont présentés dans le tableau 4. Les scores présentés correspondent à la moyenne des scores ROUGE obtenus sur les deux "blocs" qui composent un résumé.

5.6 Résultats

La méthode SenWE-KP permet d'obtenir des résultats légèrement meilleurs en terme de scores Rouge-2 et Rouge-3. Étant donné la taille de notre corpus et la faible différence de score entre les deux méthodes, on peut considérer que leurs résultats ROUGE sont comparables. Notre méthode semble donc bien capturer les aspects comparatifs entre deux jeux de documents. Les figures 3 et 4 présentent les résumés obtenus pour le sujet "Tremblement de terre Chili / Haïti 2010" respectivement avec la *baseline* WordNet-ILP et notre méthode SenWE-KP.

Néanmoins, en étudiant manuellement les résumés générés, on peut observer que ceux-ci sont en général pollués par une certaine redondance. La méthode WordNet-ILP pallie ce problème en utilisant une optimisation sur une somme de concepts et de paires de concepts. Étant donné que chaque concept et chaque paire n'est compté qu'une seule fois, le meilleur résumé doit limiter la redondance. Dans notre méthode, même si nous ne comptons qu'une seule fois chaque score tf-idf dans la partie représentativité de la fonction objectif, nous ne pouvons pas réduire les similitudes aussi facilement.

Cependant, l'utilisation d'une granularité au niveau de la phrase améliore considérablement le temps de traitement, de plus d'un facteur 10, la rendant utilisable en application réelle ; la table 4 montre que le calcul des résumés pour l'ensemble des 10 sujets de notre corpus d'évaluation prend plus de 5h pour WordNet-ILP contre 22 minutes pour notre méthode. Le temps de traitement relativement long pour la méthode WordNet provient des traitements réalisés au niveau des concepts. D'une part, le calcul des distances de similarité entre concepts dans la taxonomie WordNet est coûteux. D'autre part, le solveur de programmation linéaire en nombre entiers possède un très grand nombre de contraintes,

Résumé A (Chili) : The Geological Survey said that another earthquake on Saturday, a 6.3-magnitude quake in northern Argentina, was unrelated. The magnitude-8.8 earthquake struck at 3 :34 AM at a depth of 35km. It also recorded at least eight aftershocks, the largest of 6.9 magnitude at 0801 GMT. The earthquake struck at 0634 GMT, 115km (70 miles) north-east of the city of Concepcion and 325km south-west of the capital Santiago. This was in direct contrast to Haiti, which was unprepared for the Jan. 12 earthquake. The quake was followed by 76 aftershocks of 4.9 magnitude or greater. The damage from Chile's earthquake was widespread.
Résumé B (Haïti) : The tremor hit at 1653 (2153 GMT) on Tuesday, the US Geological Survey said. A massive 7.0-magnitude earthquake has struck the Caribbean nation of Haiti. The earthquake hit at 4 :53 PM some 15 miles (25 km) southwest of the Haitian capital of Port-au-Prince. Most severely affected was Haiti, occupying the western third of the island. At least 10 aftershocks followed, including two in the magnitude 5 range, the USGS reported. The hospital in Petionville, a well to do neighbourhood, home to diplomats and expatriates, was wrecked. In addition, less than one-third of the population was steadily employed.

FIGURE 4 – Résumé comparatif issu de la méthode SenWE-KP pour le sujet "Tremblement de terre Chili / Haiti 2010".

dû à la nécessité d'encoder les similarités entre les concepts présents dans les jeux de documents différents.

Méthodes	R2	R3	Temps d'exécution
WordNet-ILP	0.08704	0.05207	19335s
SenWE-KP	0.09006	0.05410	1352s

TABLE 4 – Scores ROUGE et temps d'exécution pour chaque méthode.

6 Conclusion et perspectives

Dans cet article, nous présentons une nouvelle méthode pour le résumé comparatif de deux jeux de documents traitant de sujets comparables. Nous utilisons les plongements de mots et la mesure de distance sémantique *Word Mover Distance* au niveau de la phrase afin de découvrir celles les plus susceptibles d'être comparatives. Nous comparons notre méthode à une méthode de l'état de l'art décrite par Huang *et al.* (2014), qui utilise WordNet pour repérer les phrases comparatives. Nos résultats confirment que l'utilisation des plongements de mots et de la *Word Mover Distance* au niveau de la phrase permet d'obtenir des résultats comparables à l'utilisation de similarités dérivées de WordNet. Nous améliorons en revanche considérablement le temps de traitement par rapport à la méthode de Huang *et al.* (2014).

Nous travaillons à l'extension de nos baselines, en y ajoutant la méthode de Campr & Ježek (2013). Une évaluation manuelle est également prévue afin de confirmer les résultats obtenus avec la méthode ROUGE.

Dans la suite de nos recherches, nous devons travailler à l'élimination de la redondance au sein du résumé comparatif, par exemple en supprimant les paires de phrases trop similaires à des phrases déjà présentes dans le résumé ou en intégrant une mesure de la redondance dans notre fonction objectif. Nous avons également constaté l'importance des entités nommées dans la détection des phrases

comparables. En effet, elles sont souvent l'élément de différenciation au sein de phrases qui traitent du même thème. Nous souhaitons donc incorporer des traitements sur les entités nommées afin de mieux identifier les phrases comparables.

Les avancées dans la recherche sur les plongements de mots permettent aujourd'hui d'obtenir des plongements de mots multilingues (Conneau *et al.*, 2017). Nous souhaitons appliquer notre méthode au résumé comparatif multilingue ou translingue, donc en travaillant sur des jeux de documents dans des langues différentes. Cela permettrait d'étendre la portée de notre travail. L'utilisation de plongements de mots multilingues nous permettrait d'appliquer directement notre méthode au résumé comparatif multilingue ou crosslingue.

Remerciements

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-16-CE38-0008 (projet ANR JCJC ASADERA).

Références

ANTTILA R. (1972). *An introduction to historical and comparative linguistics*. Macmillan New York.

BAO S., LI R., YU Y. & CAO Y. (2008). Competitor mining with the web. *IEEE Transactions on Knowledge and Data Engineering*, **20**(10), 1297–1310.

CAMPR M. & JEŽEK K. (2013). Topic models for comparative summarization. In *International Conference on Text, Speech and Dialogue*, p. 568–574 : Springer.

CONNEAU A., LAMPLE G., RANZATO M., DENOYER L. & JÉGOU H. (2017). Word translation without parallel data. In *International Conference on Learning Representations*. arXiv preprint : 1710.04087.

GAMBHIR M. & GUPTA V. (2017). Recent automatic text summarization techniques : A survey. *Artif. Intell. Rev.*, **47**(1), 1–66. DOI : 10.1007/s10462-016-9475-9.

GILLICK D. & FAVRE B. (2009). A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, ILP '09, p. 10–18, Stroudsburg, PA, USA : 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*, p. 128–137.

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.

HUANG X., WAN X. & XIAO J. (2014). Comparative news summarization using concept-based optimization. *Knowledge and information systems*, **38**, 691–716.

JAIN A. & PANTEL P. (2011). How do they compare ? automatic identification of comparable entities on the web. In *2011 IEEE International Conference on Information Reuse & Integration*, p. 228–233 : IEEE.

JINDAL N. & LIU B. (2006). 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.

KENNEDY C. (2007). Modes of comparison. In *Proceedings from the annual meeting of the Chicago Linguistic Society*, volume 43.1, p. 141–165 : Chicago Linguistic Society.

KIM H. D. & ZHAI C. (2009). Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM conference on Information and knowledge management*, p. 385–394 : ACM.

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

LERNER J.-Y. & PINKAL M. (2003). Comparatives and nested quantification. In J. GUTIÉRREZ-REXACH, Éd., *Semantics : critical concepts in linguistics. Vol. V : Operators and sentence types*, chapitre 68, p. 70–87. Routledge.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics. Anthologie ACL : W04-1013.

LIU J., WAGNER E. & BIRNBAUM L. (2007). Compare&contrast : using the web to discover comparable cases for news stories. In *Proceedings of the 16th international conference on World Wide Web*, p. 541–550.

LOUIS A. & NENKOVA A. (2013). Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, **39**(2), 267–300. DOI : 10.1162/COLI_a_00123.

MCDONALD R. (2007). A study of global inference algorithms in multi-document summarization. In *European Conference on Information Retrieval*, p. 557–564 : Springer.

NAVIGLI R. & PONZETTO S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**, 217–250.

NENKOVA A. & PASSONNEAU R. (2004). Evaluating content selection in summarization : The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics : HLT-NAACL 2004*, p. 145–152, Boston, Massachusetts, USA : Association for Computational Linguistics.

PAUL M. J., ZHAI C. & GIRJU R. (2010). Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, p. 66–76 : Association for Computational Linguistics.

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.

PISINGER D., RASMUSSEN A. & SANDVIK R. (2007). Solution of large-sized quadratic knapsack problems through aggressive reduction. *I N F O R M S Journal on Computing*, **19**(2), 280–290. DOI : 10.1287/ijoc.1050.0172.

PRAWER S. S. (1973). *Comparative literary studies : an introduction*. Duckworth London.

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 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.

WANG D., ZHU S., LI T. & GONG Y. (2012). Comparative document summarization via discriminative sentence selection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **6**, 12.

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.