Similarité textuelle pour l'association de documents journalistiques

Delphine Charlet* — **Géraldine Damnati***

* *Orange Labs, Lannion, France*
{delphine.charlet,geraldine.damnati}@orange.com

RÉSUMÉ. *Cet article étudie l'association de documents journalistiques issus de la presse en ligne et de journaux télévisés, en utilisant des similarités sémantiques textuelles. Les associations de documents sont étudiées dans des configurations intramedia et intermedia. Les expériences menées montrent que les métriques de similarité sémantique qui s'avéraient efficaces dans le contexte de similarité entre questions posées sur un forum sont également efficaces pour l'association de documents, quelle que soit la configuration d'association média. L'influence de la longueur des documents requêtes et cibles est étudiée de façon approfondie et montre des comportements contrastés des métriques selon la longueur.*

ABSTRACT. *This article explores the linking of written and audiovisual news, based on the use of semantic textual similarity metrics. It presents a comprehensive study of different linking approaches with various configurations of intermedia or intramedia association. It is shown that textual similarity metrics that have proved to perform very well in the context of community question answering can provide efficient news linking metrics, whatever the media association configuration. The influence of document length and request length is also explored for several similarity metrics. The results highlight contrasted behaviors regarding request and document lengths.*

MOTS-CLÉS : *association de documents, similarité sémantique textuelle, relations sémantiques*

KEYWORDS: *documents linking, semantic textual similarity*

1. Introduction

L'association de documents d'information est un sujet qui recouvre différents enjeux, allant de l'aide aux professionnels du journalisme afin de parcourir plus efficacement la couverture médiatique d'un événement, jusqu'à la conception d'outils grand public de comparaison de sources d'information. L'association est ainsi un préalable pour la mise en oeuvre d'outils d'exploration d'informations, e.g. (Bois *et al.*, 2017b) ou comme point d'entrée pour la conception de moteurs de recherche efficace permettant de trouver des informations en lien avec une information consultée. Nous nous proposons dans cette étude de nous concentrer sur ce dernier cas d'usage et nous établissons une comparaison entre différentes mesures de similarité dans diverses configurations de Recherche d'Information.

L'association multi-modale nécessite de définir des mesures de similarité sémantiques entre des documents de différente nature. La détermination d'hyperliens video est une tâche à part entière dans les campagnes d'évaluation MediaEval (Eskevich *et al.*, 2014) ou TRECVid (voir par exemple (Bois *et al.*, 2017c)). L'objectif est alors d'être capable de lier une ancre (un extrait de programme de la BBC sélectionné par des experts comme étant un segment d'intérêt) à d'autres segments définis comme étant des cibles, à rechercher parmi 2700 heures de programmes. Cette tâche, de façon analogue aux tâches de similarité sémantique textuelle, s'applique à des données homogènes : l'objectif est de lier un fragment à un autre fragment issu de la même source. D'autres travaux s'attachent à traiter de l'association entre des sources hétérogènes mais plutôt avec un point de vue d'alignement (*e.g.* des livres et des films (Zhu *et al.*, 2015) ou des enregistrements de conférences et les articles scientifiques associés (Mougard *et al.*, 2015)).

Dans le domaine journalistique, plusieurs études ont porté sur l'association d'articles de presse avec d'autres sources d'information. (Aker *et al.*, 2015) étudient l'association entre articles de presse écrite et de commentaires issus du site de *The Guardian*(Das *et al.*, 2014). L'association entre articles de presse et tweets a également été étudiée (Guo *et al.*, 2013). Concernant l'association entre presse écrite et presse audiovisuelle, (Henzinger *et al.*, 2005) propose d'enrichir un journal télévisé par des articles de la presse en ligne, collectés par des requêtes générées automatiquement à partir d'extraits de sous-titres. Plus récemment, (Bois *et al.*, 2017a) s'attachent à construire des représentations en graphes pour l'exploration (*News browsing*). Les auteurs ont collecté sur une période de trois semaines un corpus de documents en français incluant des articles de presse écrite, des vidéos et des podcasts de radio. Récemment le corpus *FrNewsLink* (Camelin *et al.*, 2018) a été rendu public, permettant de traiter plusieurs tâches d'association multi-modales, avec des données hétérogènes issues de sources variées et de tailles différentes. Nous nous proposons d'étudier l'influence de la mesure de similarité sémantique retenue sur la tâche qui consiste à retrouver des informations en lien avec une information considérée comme requête.

La similarité sémantique de textes a fait l'objet de plusieurs défis de recherche récents, e.g. (Cer *et al.*, 2017) pour le calcul de similarité entre phrases, ou (Nakov

et al., 2017) pour la recherche de questions similaires sur un forum, à partir d'une question requête. Dans ces défis, de nombreuses approches de calcul de similarité sémantique ont été proposées, principalement en mode supervisé, puisque des données d'apprentissage qualifiant la similarité entre paires de textes étaient disponibles. Néanmoins, la plupart de ces systèmes étaient en fait des combinaisons supervisées de mesures de similarité non-supervisées. Parmi ces mesures non supervisées, l'une d'elles s'est avérée particulièrement performante, dans le contexte du défi sur l'association de questions dans les forums. Il s'agit de la mesure `soft-cosine`, qui tire profit de relations sémantiques entre mots issues des plongements de mots, dans un formalisme qui généralise la mesure de similarité en cosinus entre sacs de mots. (Charlet et Damnati, 2017).

Dans cet article, nous explorons le potentiel de cette mesure de similarité, pour associer des données hétérogènes telles que des articles de presse en ligne et des extraits de journaux télévisés. L'utilisation de plongements de mots dans un contexte de Recherche d'Information a fait l'objet d'études récentes (Craswell *et al.*, 2017), que ce soit dans un contexte d'expansion de requêtes (Diaz *et al.*, 2016) (Amer *et al.*, 2016) ou de mesures de similarité entre requêtes et documents (Roy *et al.*, 2016). C'est sur ce dernier point que se situe notre contribution.

La section 2 présente le corpus public utilisé *FrNewsLink* et les tâches d'association étudiées. La section 3 présente les mesures de similarité étudiées qui sont ensuite évaluées et discutées dans la section 4.

2. Corpus et tâches étudiées

2.1. *Corpus multimédia*

Nous utilisons le corpus *FrNewsLink* récemment rendu public (Camelin *et al.*, 2018). Ce corpus est constitué de transcriptions automatiques de journaux télévisés (JT), et de textes d'articles de presse en ligne collectés pendant la même période. Le corpus *FrNewsLink* contient 112 JT issus de 8 chaînes différentes, enregistrées durant 2 périodes, en 2014 et 2015. Des annotations manuelles de segmentation thématique ont été faites pour chaque JT, ainsi le corpus contient des segments mono-thématiques de transcriptions automatiques de JTs.

Dans cet article, nous utilisons l'ensemble des JTs enregistrés durant la 7^{eme} semaine de 2014, qui contient 992 segments mono-thématiques issus de 86 JTs de 8 chaînes différentes.

Un ensemble d'environ 25000 articles de presse en ligne publiés durant cette période a également été collecté. Des annotations manuelles ont été faites, qui associent les articles et les segments monothématiques des JTs. Un article de presse en ligne a été associé à un segment de JTs si le titre de cet article pouvait être considéré comme un titre acceptable pour le segment monothématique du JT.

2.2. *Association inter-média*

Ce corpus multimédia annoté nous permet d'évaluer différentes tâches d'association inter-média. Ainsi, on peut considérer un segment de JT comme une requête, et rechercher tous les articles de presse en ligne associés, parmi les articles du même jour. Réciproquement, la requête peut être consituée d'un article de presse en ligne, et on recherche alors parmi les segments de JT du même jour, les segments qui peuvent lui être associés. Le tableau 1 décrit le corpus W07_14 (qui correspond à la période de la 7^{eme} semaine de 2014), pour l'association inter-média.

# segments de JTs	992
# segments de JTs avec au moins un article associé	707
# nombre moyen d'articles associés par segment ayant au moins un article associé	11.1
# articles de presse	5024
# article de presse avec au moins un segment de JT associé	1784
# nombre moyen de segments de JTs associés par article ayant au moins un article associé	4.4
# paires associées inter-média (segments de JTs associés à un article de presse du même jour)	7830
# paires potentielles (segment de JTs $\times$ articles de presse du même jour)	734 113
pourcentage de paires associées parmi les paires potentielles	1.1%

Tableau 1 – statistiques descriptives de W07_2014 pour l'association inter-média

2.3. *Association intra-média*

À partir du corpus $FrNewsLinks$ annoté en association inter-média, nous pouvons construire, par supervision indirecte, 2 corpus permettant l'évaluation de l'association intra-média.

2.3.1. *Association de segments de JTs*

2 segments de JTs sont considérés comme associés si ils sont associés à au moins un article de presse en commun. Si 2 segments de JTs sont associés à des articles de presse, sans avoir d'articles associés en commun, on considère que ces 2 segments ne sont pas associés. On ne peut pas conclure quant à l'existence d'association entre 2 segments qui ne sont associés à aucun article de presse. En effet, ils pourraient être associés, sur un sujet qui n'est pas traité dans le corpus d'articles de presse en ligne. C'est pourquoi, dans le contexte de l'évaluation de l'association de segments de JTs, nous restreignons le corpus au sous-ensemble de segments ayant au moins un article associé.

Le tableau 2 présente des statistiques décrivant le corpus pour cette tâche.

# segments de JTs (*avec au moins un article associé)*	707
# segments de JTs associés à au moins un autre segment de JTs	604
# nombre moyen de segments associés par segment ayant au moins un segment associé	11.3
# paires de segments de JTs du même jour associés entre eux	6844
# paires potentielles de segments de JTs du même jour	76444
pourcentage de paires associées parmi les paires potentielles	9.0%

Tableau 2 – statistiques descriptives de W07_14 pour l'association de segments de JTs

On peut noter que parmi les 707 segments de JTs ayant au moins un article de presse associé, 85% d'entre eux (604) ont également un segment de JTs associés. Seuls 15% des segments associés à des articles de presse n'ont pas de liens à d'autres segments. Cela signifie que lorsqu'un sujet de télévision est également traité dans la presse en ligne, il est très probable que ce sujet soit également traité dans d'autres JTs du même jour.

2.3.2. *Association d'articles de presse*

Réciproquement, nous pouvons appliquer la même approche pour la constitution d'un corpus d'articles de presse associés. 2 articles de presses sont considérés comme associés entre eux si ils sont associés à au moins un segment de JT en commun. Si 2 articles de presse, associés à des segments de JT, n'ont pas de segments de JTs en commun, on considère qu'ils ne sont pas associés. On ne peut pas conclure quant à l'existence d'une association entre 2 articles n'ayant aucun segments de JTs associés (ils peuvent être néanmoins associés, sur un sujet non traité par les JTs). C'est pourquoi dans le contexte de l'évaluation de l'association d'articles de presse, nous restreignons le corpus au sous-ensemble d'articles ayant au moins un segment de JT associé.

Le tableau 3 présente des statistiques décrivant le corpus extrait pour l'évaluation de l'association d'articles de presse.

On peut noter que parmi les 1784 articles de presse qui sont associés à au moins un segment de JTs, 97% d'entre eux (1734) sont aussi associés à un autre article de presse. Cela signifie que dès lors qu'un sujet de la presse en ligne est également traité par la télévision, il est extrêmement probable que ce sujet soit traité par d'autres articles de presse en ligne. Cela confirme et amplifie le constat que nous avions fait dans la section précédente sur l'association de segments de JTs. Ainsi, l'existence pour un sujet spécifique d'une association intermédia entre la télévision et la presse en ligne implique une forte probabilité, pour ce sujet, d'être repris plusieurs fois, au sein de chaque média.

# articles de presse (*avec au moins un segment de JT associé)*	1784
# articles de presse avec au moins un article associé	1734
# nombre moyen d'articles associés par article ayant au moins un article associé	20.8
# paires d'articles du même jour associés	36126
# paires potentielles d'articles du même jour	482132
pourcentage de paires associées parmi les paires d'articles du même jour	7.5%

Tableau 3 – statistiques descriptive de W07_14 pour l'association d'articles de presse

3. Mesures de similarité textuelles

Nous étudions le comportement de la mesure récemment proposée dans (Charlet et Damnati, 2017), comparativement à 2 mesures choisies comme *baseline*. D'une part, le cosinus entre sacs de mots, reste une référence solide dans la famille des mesures basées sur les représentations creuses (e.g. Jaccard et ses variantes). D'autre part, l'autre mesure, basée sur une représentation moyenne de plongements de mots, est également une référence habituelle, dès lors qu'on explore les représentations de documents par plongements.

3.1. *Prétraitement et baseline*

Les textes sont lemmatisés et seuls les lemmes des noms, verbes et adjectifs sont sélectionnés. Les lemmes sont associés à des poids Okapi TF-IDF$_{BM25}$, estimés sur tout le corpus des articles de presse de la semaine considérée. Ainsi, les textes sont représentés par un sac de lemmes pondérés. Comme mesure de similarité *baseline*, on calcule un cosinus entre les vecteurs X et Y représentant les textes T_X et T_Y :

$$cos(X,Y) = \frac{X^t.Y}{\sqrt{X^t.X}\sqrt{Y^t.Y}} \text{avec } X^t.Y = \sum_{i=1}^{n} x_i y_i \quad [1]$$

3.2. *similarité* `soft-cosinus`

Lorsque les textes T_X et T_Y n'ont aucun mot en commun (i.e. il n'y a aucun indice i pour lequel les poids x_i et y_i sont non nuls) la similarité en *cosinus* est nulle. Cependant, même sans aucun mot en commun, des textes peuvent être liés sémantiquement, si ils contiennent des mots liés sémantiquement. C'est pourquoi des auteurs (Sidorov *et al.*, 2014) (Charlet et Damnati, 2017) ont proposé de prendre en compte les relations

entre mots, en introduisant dans la formule de la similarité en cosinus une matrice M de relations entre mots, comme dans l'équation 2.

$$cos_M(X,Y) = \frac{X^t.M.Y}{\sqrt{X^t.M.X}\sqrt{Y^t.M.Y}} \quad [2]$$

$$X^t.M.Y = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i m_{i,j} y_j \quad [3]$$

où M est une matrice dont l'élément $m_{i,j}$ exprime un lien entre le mot i et le mot j Avec une telle métrique, la similarité entre deux textes est non-nulle dès lors que les textes ont des mots liés entre eux, même s'ils n'ont aucun mot en commun. La présence de la matrice M au dénominateur est un facteur de normalisation qui assure la réflexivité de la mesure de similarité (la similarité d'un texte par rapport à lui-même est égale à 1).

Ici, M est basée sur une similarité entre plongements des mots. Cette mesure s'est avérée très efficace dans SemEval2017, pour mesurer des similarités entre questions posées sur un forum (Charlet et Damnati, 2017). Comme dans le papier sus-cité, l'élément $m_{i,j}$ est calculé de la façon suivante :

$$m_{i,j} = max(0, cos(v_i, v_j))^2 \quad [4]$$

où v_i et v_j sont les plongements des mots i et j, estimé par l'outil word2vec (Mikolov *et al.*, 2013) sur le corpus complet des articles de la semaine.

3.3. *Plongement de textes*

Une représentation très simple mais efficace d'un texte consiste en la simple moyenne des plongements des mots qui le composent. Cette représentation a été utilisée par exemple par de nombreux participants du dernier défi SemEval pour la tâche sur les forums (Nakov *et al.*, 2017). Pondérer la contribution de chaque mot dans la moyenne des plongements s'est avéré particulièrement intéressant dans divers travaux dont (Arora *et al.*, 2017) et fournit des plongements de textes très simples à calculer, mais compétitifs par rapport à d'autres méthodes plus sophistiquées. Ainsi, en considérant x_i le poids du mot i et $v_{i,k}$ le k-ème composant du mot i dans l'espace de plongements, le k-ème composant du vecteur $\tilde{X}$ qui représente le texte T_X est :

$$\tilde{X}_k = \frac{1}{\sum_i x_i} \sum_i x_i v_{i,k}$$

La mesure de similarité `wavg-w2v`(X,Y) entre T_X et T_Y est alors calculée comme le cosinus entre $\tilde{X}$ et $\tilde{Y}$. Ici, nous utilisons les mêmes poids Okapi que dans les métriques `cosinus` et `cosinus` basées sur les sacs de mots.

4. Expériences

4.1. *Protocole et mesures d'évaluation*

Nous adoptons un protocole général, quel que soit le type de textes à associer (segments de JTs ou articles de presse en ligne) La tâche qui consiste à retrouver, pour un texte-requête donné, les textes associés, est évaluée en *Mean Average Precision* (MAP). Pour une requête, les similarités entre la requête et tous les textes potentiels du même jour sont calculées, et les textes sont ordonnés par similarité textuelle décroissante. MAP@10 est utilisée pour évaluer la pertinence du tri des 10 textes les plus similaires.

En complément de cette tâche, on considère également la tâche de détection des paires correctement associées parmi toutes les paires potentielles. Les similarités sont calculées entre toutes les paires potentielles de textes, et celles dont la similarité est supérieure à un seuil donné sont considérées comme associées. Pour cet ensemble de paires détectées automatiquement comme associées, on peut calculer un taux de précision (combien sont effectivement correctement associées ?), de rappel (parmi les paires associées dans la vérité-terrain), ainsi que leur moyenne harmonique la F-measure.

MAP et F-mesure évaluent des propriétés différentes : la MAP traduit la capacité de la mesure de similarité à donner aux paires associées une meilleure valeur qu'aux paires non-associées, sans notion de seuil de décision. La F-mesure traduit en plus la capacité de la métrique à permettre un seuil de décision sur les valeurs de similarité, pour décider si une paire contient des textes associés ou non, ce seuil étant commun à toutes les requêtes. Au-delà de la capacité à ordonner, une bonne F-mesure traduit le fait que les mesures de similarité entre paires de textes distinctes soient comparables. Dans nos tableaux de résultats, le seuil de décision est fixé *a posteriori*, de façon à obtenir la F-mesure maximale (Fmax).

Les articles de presse étant composés d'un titre (la premier ligne du texte) et d'un corps (le reste du texte), des expériences contrastives sont systématiquement menées, en considérant, pour représenter l'article dans les calculs de similarités, soit uniquement son titre, soit l'article entier. En effet, on s'attend à ce que la longueur des textes disponibles pour le cacul de la mesure de similarité, ait une influence signficative sur les performances.

La longueur moyenne des segments de JTs (en termes de lemmes différents sélectionnés) est de 42.1 mots. Pour les articles de presse en ligne, la longueur moyenne du titre est 6.6 mots, alors que la longueur moyenne de l'article complet est de 120.8 mots.

requête : segment de JT		MAP@10	Fmax
cible : segments de JT	`cosinus`	0.896	61.5
	`soft-cosinus`	**0.932**	66.8
	`wavg-w2v`	**0.930**	**68.0**

Tableau 4 – association intra-media de segments de JT

4.2. *Résultats*

On considère tout d'abord les expériences où la requête est le segment de JT. Le tableau 4 présente les résultats de l'association intra-média de segments de JT. Il s'agit de retrouver les segments de JT qui traitent du même sujet que le segment requête. `soft-cosinus` et `wavg-w2v`, qui sont les métriques qui exploitent les plongements de mots, ont des performances équivalentes (avec un avantage pour `wavg-w2v` pour la Fmax) et significativement[1] meilleures que le `cosinus` de sacs de mots. On peut remarquer que, si les performances MAP sont très bonnes, la Fmax qui implique un seuil de décision global, n'est pas aussi bonne.

Ensuite, le tableau 5 présente les résultats de l'association d'articles de presse à des segments de JT, avec 2 variantes : dans l'une, les articles sont uniquement représentés par leur titre, tandis que dans l'autre, on considère tout l'article. Quand il s'agit d'associer des segments de JTs à des textes très courts (titres d'articles), les performances sont moins bonnes que celles obtenues quand on considère les articles entiers. Il est intéressant de constater que si les métriques utilisant les plongements de mots sont bien plus performantes que le cosinus des sacs de mots pour l'association avec les titres, ce n'est pas le cas pour l'association avec les textes plus longs que sont les articles entiers. Dans ce cas, c'est le cosinus entre sacs de mots qui donne les meilleures performances en MAP. Ces résultats sont cohérents avec le fait que l'avantage des métriques utilisant les plongements de mots est de pouvoir calculer des similarités entre textes qui n'ont pas de mots en commun. Plus les textes sont courts, plus il est probable que des textes liés n'aient pas ou peu de mots en commun. Quand les textes sont longs, il est probable que des textes liés aient beaucoup de mots en commun. Dans ce cas, utiliser les proximités des plongements de mots dans les métriques peut introduire du bruit et perturber les mesures. Cependant, du point de vue de la Fmax, la mesure `soft-cosinus` est la métrique qui donne les meilleures performances quelque soit la taille du document cible. Cela suggère que cette métrique est la meilleure dans sa capacité à permettre un seuil de décision global pour la détection de paires associées.

Le tableau 6 présente les résultats obtenus pour la tâche symétrique de la tâche précédente : la requête est désormais l'article de presse, pour lequel on cherche les segments de JT associés. Les variantes où l'on considère soit le titre de l'article soit

1. le test de significativité est le t-test, en considérant comme observation les performances de *average precision* @10 par requête, la p-value obtenue est inférieure à 0.01

requête : segment de JT		MAP@10	Fmax
cible : titre d'articles de presse	`cosinus`	0.680	53.7
	`soft-cosinus`	**0.750**	**66.1**
	`wavg-w2v`	0.743	65.1
cible : articles entiers	`cosinus`	**0.834**	73.5
	`soft-cosinus`	0.820	**74.5**
	`wavg-w2v`	0.807	72.1

Tableau 5 – Associer des articles de presse à des segments de JTs

requête : titre d'article de presse		MAP@10	Fmax
cible : segments de JT	`cosinus`	0.761	53.7
	`soft-cosinus`	0.**889**	**66.1**
	`wavg-w2v`	**0.887**	65.1
requête : article entier		MAP@10	Fmax
cible : segments de JT	`cosinus`	0.917	73.5
	`soft-cosinus`	**0.923**	**74.5**
	`wavg-w2v`	0.896	72.1

Tableau 6 – associer des segments de JTs à des articles de presse

l'article complet sont étudiées. Ici encore, on observe que les performances obtenues en utilisant l'article complet sont meilleures que celles obtenues en utilisant le titre. C'est également la métrique `cosinus` qui est la plus sensible à la longueur de la requête : sa MAP est égale à 0.761 quand la requête est le titre et atteint 0.917 quand la requête est l'article complet. A l'opposé, la métrique `wavg-w2v` ne tire pas beaucoup profit de l'augmentation de la longueur de la requête : ainsi, sa MAP est égale à 0.887 quand la requête est le titre et n'atteind que 0.896 pour l'article complet. La métrique `soft-cosine` est la plus performante, quelque soit la longueur de la requête.

Enfin, le tableau 7 présente les résultats obtenus pour l'association intra-média d'articles de presse, avec toutes les variantes possibles, que la requête ou la cible soient constituées du titre uniquement ou de l'article complet.

Quand il s'agit d'associer des textes très courts ensemble (titre/titre), la mesure `wavg-w2v` est significativement la plus performante. La mesure `soft-cosinus`, qui utilise également les plongements de mots, obtient des performances meilleures que la mesure `cosinus`, mais moins bonnes cependant que `wavg-w2v`. Quand il s'agit d'associer de longs textes (article complet/article complet), `cosinus` et `soft-cosinus` obtiennent la meilleure MAP. On peut noter que, pour une MAP équivalente, la mesure `cosinus` obtient une bien moins bonne Fmax que les autres mesures. Par exemple, dans le cas de l'association (article complet/titre), `cosinus` et `wavg-w2v`

requête : titre d'article		MAP@10	Fmax
cible : titre d'article	`cosinus`	0.807	56.2
	`soft-cosinus`	0.865	58.6
	`wavg-w2v`	**0.910**	**75.4**
cible : article complet	`cosinus`	0.907	67.7
	`soft-cosinus`	**0.941**	**80.3**
	`wavg-w2v`	0.933	79.3
requête : article complet		MAP@10	Fmax
cible : titre d'article	`cosinus`	0.918	67.7
	`soft-cosinus`	**0.933**	**80.3**
	`wavg-w2v`	0.911	79.3
cible : article complet	`cosinus`	**0.940**	78.9
	`soft-cosinus`	**0.939**	**83.7**
	`wavg-w2v`	0.927	82.3

Tableau 7 – association intra-media d'articles de presse en ligne

obtiennent une MAP autour de 0.91 mais une Fmax respectivement de 67.7 et 79.3. De même, dans le cas de l'association (article complet/article complet), `cosinus` et `soft-cosinus` obtiennent une MAP autour de 0.94 mais une Fmax respectivement de 78.9 et 83.7. Cela signifie que la mesure `cosinus` entre sacs de mots produit des scores qui sont pertinents pour le tri, mais pas aussi pertinents que les scores des autres métriques, pour la comparaison à un seuil de décision global. Enfin, quand il s'agit d'associer des contenus de longueur très différentes (titre/article complet ou article complet/titre), la mesure la plus performante est la mesure `soft-cosinus`

4.3. *Expériences complémentaires d'associations intra-media*

Les expériences précédentes ont mis en évidence la sensibilité des différentes mesures de similarité à la longueur des textes à associer. Nous approfondissons ce point dans cette section, dans le cas de l'association intra-media d'articles de presse.

Nous adoptons le protocole suivant : nous faisons varier la taille du texte considéré, que ce soit pour la requête ou la cible, en sélectionnant un nombre de phrases variables à partir du titre. Ainsi, pour un nombre k fixé, l'article est représenté par son titre et ses k premières phrases, si l'article contient plus de k phrases, sinon l'article complet. Dans le sous-ensemble d'articles de presses que nous utilisons pour l'association intra-média, 25% des articles ont moins de 10 phrases, 75% ont moins de 20 phrases et 99% ont moins de 100 phrases. Le dernier pourcent des articles de plus de 100 phrases est en fait composé d'articles bruités, contenant un long fil de discussion associé. Nous faisons varier k de 0 (l'article n'est représenté que par son titre) à 100 (l'article est

représenté par son titre et ses 100 premières phrases au maximum, ce qui correspond à l'article complet dans 99% des cas)

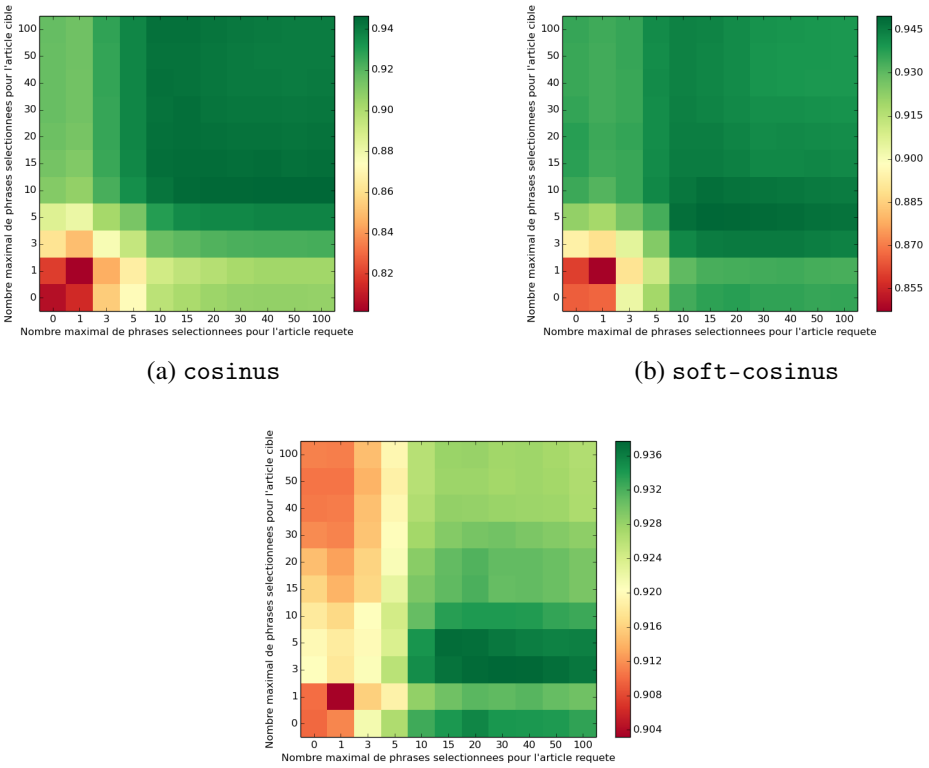
Les figures 1 représentent la carte des performances MAP@10 obtenues en faisant varier k pour la requête et pour la cible. On obtient ainsi une matrice de performances pour chacune des mesures de similarité considérées. L'observation de ces figures nous permet de constater que si les métriques `cosinus` et `soft-cosine` présentent des profils d'évolution des performances relativement proches, ce n'est pas le cas de la mesure `wavg-w2v`. D'une part, la carte des performances semble plus symétrique pour les métriques `cosinus` et `soft-cosine`, tandis qu'elle est clairement assymétrique pour `wavg-w2v`. La différence de longueur entre requête et cible semble jouer un rôle plus important dans cette dernière métrique. D'autre part, pour les 2 premières mesures, on observe globalement, pour une longueur de requête fixée, une augmentation des performances à mesure que la taille des articles cible augmente, ce qui n'est pas le cas de la mesure `wavg-w2v`, où pour une requête très courte (titre, titre+1 ou 2 phrases), les performances n'augmentent pas en fonction de la longueur des articles cibles. On constate également que les performances maximales ne sont pas obtenues en utilisant la quantité maximale de texte disponible pour la requête et la cible, quelle que soit la mesure de similarité. Ainsi, un optimum semble atteint, pour des requêtes longues (en considérant le titre et au moins les 10 ou 15 premières phrases) et des cibles d'une taille modérée (titre + 3 à 10 phrases, selon les métriques).

Pour pouvoir comparer plus facilement les mesures de similarité, la figure 2 représente les performances MAP en fonction de la longueur de la portion de l'article sélectionné, pour 2 types de requêtes : la requête restreinte au titre de l'article, ou la requête égale à l'article complet. Dans cette figure, la longueur de la portion de l'article sélectionnée, définie en nombre des k premières phrases est traduite en nombre de mots moyens contenus dans ces k premières phrases.

On constate que, pour une requête restreinte au titre, la mesure `wavg-w2v` est la meilleure pour des documents courts et est rattrapée par la mesure `soft-cosinus` pour des articles à partir de 80 mots. Pour des requêtes correspondant à l'article complet, la mesure `wavg-w2v` est toujours la moins bonne, tandis que la mesure `soft-cosinus` est toujours la meilleure, rattrapée par la mesure `cosinus` pour les longs documents. On peut également constater que l'amplitude de l'amélioration due à l'augmentation de la taille des documents cibles est bien plus importante dans le cas des requêtes courtes (égales au titre) que des requêtes longues (égales à l'article complet).

5. Discussion

De ces différentes expériences, nous pouvons tirer quelques enseignements sur les mesures de similarité. Nous résumons les comportement des métriques dans le tableau 8 : dans ce tableau, selon la longueur des documents sources et cibles, les mesures de similarité donnant les meilleures performances sont reportées. "Court" signifie que le



(a) `cosinus`

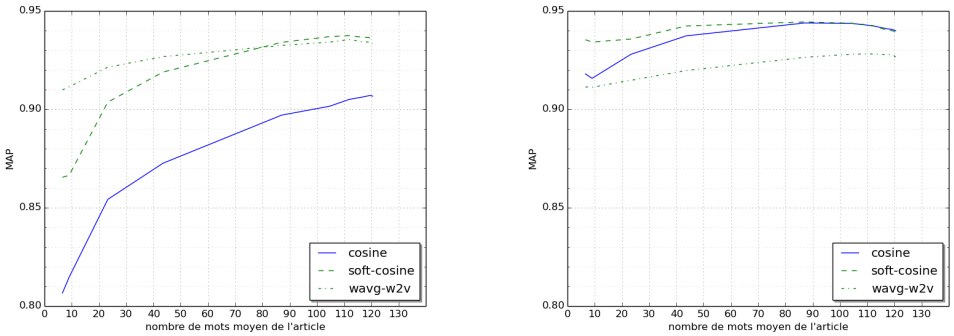
(b) `soft-cosinus`

(c) `wavg-w2v`

Figure 1 – Carte des performances MAP en fonction du nombre maximal de phrases sélectionnées pour représenter l'article

document est un titre d'environ 7 mots en moyenne, "moyen" signifie que le document est par exemple un segment de JTs d'environ 42 mots en moyenne, et "long" signifie que le document est un article complet d'environ 120 mots en moyenne.

Nous pouvons constater que les mesures utilisant les plongements de mots sont toujours meilleures ou non significativement différentes de la meilleure que le cosinus entre sacs de mots. La similarité `cosinus` entre sacs de mots est la plus sensible à la longueur des textes à associer. Elle est la plus mauvaise mesure pour les textes courts, tandis qu'elle est la meilleure, ou proche de la meilleure, pour les textes longs. `wavg-w2v` est de loin la meilleure mesure pour associer des textes courts, mais elle est sensible à la différence de longueur entre les textes à associer et n'est pas non plus très efficace pour associer des textes longs. La mesure `soft-sosinus` est la plus robuste



(a) requête : article titre (b) requête : article complet

Figure 2 – performances en fonction de la longueur des articles cibles

cible/requête	court	moyen	long
court	wavg-w2v	wavg-w2v soft-cosinus	soft-cosinus
moyen	wavg-w2v soft-cosinus	wavg-w2v soft-cosinus	soft-cosinus cosinus
long	wavg-w2v soft-cosinus	soft-cosinus cosinus	soft-cosinus cosinus

Tableau 8 – Mesures de similarité donnant les meilleures performances MAP selon la longueur des documents

au décalage de longueur de textes, et offre les meilleures performances dans toutes les configurations, à l'exception notable de l'association de textes courts.

Bien que ces mesures soient par définition symétriques, évaluées en MAP dans un contexte de RI, avec un rôle dissymétrique donné à chaque texte (un texte requête et des textes cibles), elles peuvent avoir un comportement dissymétrique, selon ce qu'on considère être la requête et la cible. Les mesures `cosinus` et `soft-cosinus` ont un comportement symétrique, ce qui n'est pas le cas de `wavg-w2v` qui, pour des requêtes courtes, ne tire pas profit de l'augmentation des textes cibles.

Enfin, si les résultats évalués en MAP sont élevés, les performances en F-mesure sont bien moins élevées. Pour obtenir de bonnes performances en F-mesure, il faut que les scores de similarité soient comparables, quelles que soient les paires, pour pouvoir être soumis à un seuil de décision global. Les mesures de similarité doivent être approfondies dans ce sens.

6. Conclusion

Dans cet article, nous avons étudié l'association de documents journalistiques issus de la presse en ligne et de journaux télévisés, basée sur la similarité textuelle entre les documents. A partir d'un corpus rendu récemment public d'association de documents inter-média, nous avons déduit des corpus d'association de documents intra-media. Nous avons ensuite étudié toutes les configurations d'association inter et intra-média possibles. Ces associations de documents ont été faites à partir de similarités textuelles et le comportement d'une mesure de similarité sémantique récemment proposée dans un contexte d'associations de questions sur un forum a été étudié de façon approfondie. L'étude systématique de l'influence de la taille des documents nous a permis de dresser un tableau contrasté des comportements des mesures de similarité à l'égard de la longueur des documents. Sur ce dernier aspect, ces travaux pourraient se prolonger en explorant la question de l'extraction (pour les documents longs) ou de la génération (pour les documents courts) du texte le plus pertinent dans chaque document pour effectuer cette association, en s'inspirant des approches de génération et d'expansion de requêtes couramment utilisées en RI.

7. Bibliographie

Aker A., Kurtic E., Hepple M., Gaizauskas R., Di Fabbrizio G., « Comment-to-article linking in the online news domain », *Proceedings of the SIGDIAL 2015 Conference*, ACL, p. 245-249, 2015.

Amer N. O., Mulhem P., Géry M., « Toward word embedding for personalized information retrieval », *Neu-IR : The SIGIR 2016 Workshop on Neural Information Retrieval*, 2016.

Arora S., Liang Y., Ma T., « A simple but tough-to-beat baseline for sentence embeddings », *Proceedings of ICLR 2017*, Toulon, France, April, 2017.

Bois R., Gravier G., Jamet É., Morin E., Robert M., Sébillot P., « Linking Multimedia Content for Efficient News Browsing », *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR 2017, Bucharest, Romania, June 6-9, 2017*, p. 301-307, 2017a.

Bois R., Gravier G., Jamet E., Morin E., Sébillot P., Robert M., « Language-based Construction of Explorable News Graphs for Journalists », *Proceedings of the 2017 EMNLP Workshop Natural Language Processing meets Journalism*, Association for Computational Linguistics, Copenhagen, Denmark, p. 31-36, September, 2017b.

Bois R., Vukotić V., Simon A.-R., Sicre R., Raymond C., Sébillot P., Gravier G., *Exploiting Multimodality in Video Hyperlinking to Improve Target Diversity*, Springer International Publishing, Cham, p. 185-197, 2017c.

Camelin N., Damnati G., Bouchekif A., Landeau A., Charlet D., Estève Y., « FrNewsLink a corpus linking TV Broadcast News Segments and Press Articles », *Proceedings of LREC 2018*, Miyazaki, Japan, May, 2018.

Cer D. M., Diab M. T., Agirre E., Lopez-Gazpio I., Specia L., « SemEval-2017 Task 1 Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation », *Proceedings of the*

11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017, p. 1-14, 2017.

Charlet D., Damnati G., « SimBow at SemEval-2017 Task 3 Soft-Cosine Semantic Similarity between Questions for Community Question Answering », *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, p. 315-319, 2017.

Craswell N., Croft W. B., Guo J., Mitra B., de Rijke M., « Report on the sigir 2016 workshop on neural information retrieval (neu-ir) », *ACM Sigir forum*, vol. 50, ACM, p. 96-103, 2017.

Das M. K., Bansal T., Bhattacharyya C., « Going beyond Corr-LDA for detecting specific comments on news & blogs », *Proceedings of the 7th ACM international conference on Web search and data mining*, ACM, p. 483-492, 2014.

Diaz F., Mitra B., Craswell N., « Query expansion with locally-trained word embeddings », *arXiv preprint arXiv :1605.07891*, 2016.

Eskevich M., Aly R., Racca D., Ordelman R., Chen S., Jones G. J., « The search and hyperlinking task at MediaEval 2014 », 2014.

Guo W., Li H., Ji H., Diab M. T., « Linking Tweets to News A Framework to Enrich Short Text Data in Social Media. », *ACL (1)*, p. 239-249, 2013.

Henzinger M., Chang B.-W., Milch B., Brin S., « Query-free news search », *World Wide Web*, vol. 8, n° 2, p. 101-126, 2005.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., « Distributed representations of words and phrases and their compositionality », *Advances in neural information processing systems*, 2013.

Mougard H., Riou M., de la Higuera C., Quiniou S., Aubert O., « The Paper or the Video Why Choose ? », *Proceedings of the 24th International Conference on World Wide Web*, ACM, p. 1019-1022, 2015.

Nakov P., Hoogeveen D., Màrquez L., Moschitti A., Mubarak H., Baldwin T., Verspoor K., « SemEval-2017 Task 3 Community Question Answering », *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Association for Computational Linguistics, Vancouver, Canada, August, 2017.

Roy D., Ganguly D., Mitra M., Jones G. J., « Representing documents and queries as sets of word embedded vectors for information retrieval », *arXiv preprint arXiv :1606.07869*, 2016.

Sidorov G., Gelbukh A. F., Gómez-Adorno H., Pinto D., « Soft Similarity and Soft Cosine Measure Similarity of Features in Vector Space Model », *Computación y Sistemas*, 2014.

Zhu Y., Kiros R., Zemel R., Salakhutdinov R., Urtasun R., Torralba A., Fidler S., « Aligning books and movies Towards story-like visual explanations by watching movies and reading books », *Proceedings of the IEEE international conference on computer vision*, p. 19-27, 2015.