Corpus d'entraînement sur les plongements de mots pour la recherche de microblogs culturels

Nayanika Dogra, Philippe Mulhem, Lorraine Goeuriot, Massih Amini-Reza

Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP [1]*, LIG, 38000 Grenoble, France*
prenom.nom@imag.fr

RÉSUMÉ. *Cet article décrit un cadre expérimental et des résultats obtenus pour la recherche de microblogs. Notre approche consiste à étudier de quelle manière l'apport de l'utilisation de plongements de mots, très populaire actuellement en recherche d'information, est dépendant de l'ensemble d'apprentissage de ces plongements. Nous étudions en particulier son utilisation pour étendre des requêtes sur des tweets culturels sur le corpus CLEF CMC 2016. Nos résultats montrent que l'utilisation de corpus spécifiques (au niveau sujet ou bien sujet+type de document) ne fournit pas forcément de meilleurs résultats.*

ABSTRACT. *We describe here an experimental framework and the results obtained on microblogs retrieval. We study the contribution one popular approach, i.e., words embeddings, depends on the learning set used to train the embeddings. We focus on query expansion for the retrieval of tweets on the CLEF CMC 2016 corpus. We find that specific corpus regarding topicality and document types does not always lead to better results.*

MOTS-CLÉS : *Plongement de mots, expansion de requêtes, microblogs*

KEYWORDS: *words embeddings, query expansion, microblogs*

DOI:10.3166/DN..-.1-15

1. Institute of Engineering Univ. Grenoble Alpes

1. Introduction

Le principe général de la recherche d'information est de retrouver les documents les plus pertinents répondant à la requête d'un utilisateur. Pour cela, un utilisateur pose une requête qui est traitée par un système de recherche d'information. En réponse, ce système fournit une liste de documents, triés par valeurs décroissantes de pertinence d'après les score calculés par le système. Dans le cas où les intersections entre les termes des requêtes et des documents sont faibles, il est nécessaire d'intégrer au processus de recherche des similarités entre les mots, afin de diminuer le silence dans les réponses et d'améliorer la pertinence des résultats. Cela est le cas par exemple pour la recherche de microblogs (publications courtes telles que les tweets). Nous proposons d'utiliser des similarités entre mots appris par des plongements de mots, et de les utiliser en étendant les requêtes des utilisateurs. Nous nous situons dans un cadre spécifique : la recherche de microblogs culturels en langue française, c'est-à-dire des tweets émis durant des festival musicaux. Les corpus d'apprentissage considérés diffèrents du point de vue sujet (général ou relatif à la musique) ainsi que du point de vue du type de texte ("classique" tiré de Wikipédia ou bien microblogs).

Plus précisément, dans cet article, nous ne proposons pas de nouvelles manières de prendre en compte les plongements de mots pour faire de l'expansion de requête, mais nous étudions l'impact d'un paramètre très important de l'apprentissage des plongements, le *corpus d'aprentissage*, sur la qualité de la recherche. Les approches d'expansions utilisées sont tirées de [Almasri et al. 2017,Almasri 2017]. Ces approches ont le mérite d'être assez simple, et donc de nous permettre d'estimer l'apport des plongements de manière directe, en évitant d'intégrer de nombreuses étapes qui pourraient "noyer" l'analyse de l'apport brut des apprentissages. Nos résultats montrent qu'il est préférable d'utiliser, dans notre cadre, des plongements appris sur le corpus utilisé par le système de RI (corpus de microblogs sur le sujet spécifique des festivals de musique), même si ce corpus n'est pas très volumineux.

Cet article suit le plan suivant. Nous commençons par décrire des travaux de l'état de l'art relatifs aux microblogs et à l'utilisation des plongements de mots en partie 2. Nous décrivons ensuite les approches que nous étudions pour l'expansion de requêtes avec des plongements en section 3. Nos expérimentations et les résultats obtenus sont rapportés en partie 4, avant de conclure.

2. État de l'art

Nous nous concentrons ici sur deux aspects : le premier est de déterminer les problèmes inhérents à la recherche de microblogs, et le second se concentre sur l'utilisation de plongements de mots pour la recherche d'information, surtout dans le cadre de l'expansion de requêtes.

2.1. La problématique des microblogs

Les microblogs sont des documents spécifiques pour différentes raisons (Jabeur *et al*., 2012 ; Gimpel *et al*., 2011 ; Twitter, 2018) :

– ils sont courts (quelques centaines de caractères);

– ils utilisent des mots-clés spécifiques (comme des *hashtags* dans Twitter) qui permettent de distinguer de manière explicite des sujets dans le microblog;

– ils peuvent contenir des liens pointant sur des informations supplémentaires (images, etc.) extérieures au microblog;

– ils utilisent des syntaxes et des vocabulaires spécifiques, comme des abréviations, des émoticons, etc.;

– ils sont souvent utilisés pour émettre des avis, des opinions (Mohammad *et al*., 2017). Ils sont donc fortement subjectifs.

Le recherche de tels documents rebat les cartes de la recherche d'information classique qui repose sur l'intersection des termes en requêtes et documents. Dès lors, il est nécessaire de se poser des questions pour tenter d'apporter des réponses à ce cadre spécifique. Une idée assez simple est de rester sur un système de recherche d'information classique, et d'utiliser des ressources externes quelconques pour améliorer l'intersection entre les termes des requêtes et des documents en étendant artificiellement le contenu. Ces ressources externes ont en fait comme objectif de réaliser une expansion des termes des documents ou des requêtes (ou des deux), afin de diminuer un éventuel silence dans les réponses.

2.2. Les plongements de mots pour la RI

Les plongements de mots, imaginés par (Bengio *et al*., 2003) ont été rendus célèbres par (Mikolov *et al*., 2013). Ils reposent sur des techniques d'apprentissage qui permettent de représenter un mot dans un espace multidimensionnel continu (le plongement) qui prend en compte les contextes d'occurrences de chaque mots du corpus d'apprentissage. Ces plongements sont clairement une avancée pour le traitement de la langue (Iacobacci *et al*., 2015), ainsi que pour la recherche d'information (augmentation des papiers à la conférence ACM SIGIR sur ce sujet (Mitra, Craswell, 2017), revue de la littérature (Zhang *et al*., 2016)). Les approches à base de plongements ont un certain nombre de paramètres, comme l'architecture d'apprentissage (*Continuous Bag of Words*, CBOW, ou bien *Skip-gram*), les hyperparamètres d'apprentissage (nombre d'itérations, taux d'apprentissage, sous-échantillonage, taux d'échantillonnage négatif, la taille des fenêtres pour les contextes des occurrences de mots, la dimension de l'espace de plongement, ...), ainsi que le corpus d'apprentissage utilisé pour générer les plongements. Dans différents cadres, comme par exemple dans le domaine biomédical (Chiu *et al*., 2016), l'impact des hyperparamètres a été bien étudié. Certains travaux, comme (Chiu *et al*., 2016), se sont intéressés à l'impact du corpus d'apprentissage, en concluant qu'il est difficile d'estimer la qualité de ce qui est appris sur un corpus quand on l'applique à un autre corpus. Notre idée ici est de se focaliser sur

les corpus pour la RI et contenant des microblogs culturels afin d'étudier dans quelle mesure ils impactent la qualité de la recherche d'information.

Dans le cadre de la recherche d'information, l'utilisation de plongements de mots est étudiée depuis quelques années. Comme le montrent (Kuzi *et al.*, 2016 ; Zhang *et al.*, 2016), on peut utiliser les plongements à différents moments du processus de recherche. Parmi les utilisations existantes des plongements de mots, comme (Kuzi *et al.*, 2016 ; Almasri *et al.*, 2016 ; Roy *et al.*, 2016), le choix de l'ensemble d'apprentissage pour les plongements est éludé : tous ces travaux utilisent le corpus de documents du système de RI. Cette question est cependant abordée dans (Diaz *et al.*, 2016) pour l'expansion de requêtes pour des documents "classiques" (c'est-à-dire pas des microblogs); les auteurs concluent qu'utiliser les corpus de RI pour l'apprentissage donne en général de meilleurs résultats. Ce résultat sert de base aux travaux rapportés ici, mais nous explorons cette problématique dans le cas de documents spécifiques, les microblogs, et d'une thématique spécifique, les festivals musicaux.

3. Expansion de requêtes

Notre approche consiste tout d'abord à fixer un cadre général dans lequel nous allons étudier l'impact des corpus d'apprentissage pour la recherche de microblogs. Comme nous l'avons vu précédemment, les plongements de mots peuvent être utilisés pour étendre les documents ou les requêtes. L'expansion des documents par de tels outils pose clairement la question de l'indépendance des termes, qui dès lors n'est plus liée à la sémantique des documents originaux, mais à un processus de construction. Pour se préserver de tels problèmes potentiels, nous utilisons l'expansion de requêtes comme cadre de comparaison.

Une fois le choix de l'utilisation d'expansion de requêtes avec les plongements fait, la question suivante est de déterminer de quelle manière utiliser ces plongements lors de l'expansion de requêtes. Comme nous l'avons vu plus haut, il est possible d'utiliser des approches relativement simples, mais efficaces, pour réaliser cette expansion, en particulier les approches développées par (Almasri *et al.*, 2016 ; Roy *et al.*, 2016), que nous présentons selon un cadre uniforme ici.

Nous commençons par décrire le résultat de l'apprentissage d'un plongement de mots comme une fonction pl_p, de paramètres p définie par:

$$\begin{aligned} pl_p : \mathcal{V} &\rightarrow \mathbb{R}^n \\ w &\mapsto pl_p(w), \end{aligned} \tag{1}$$

où $\mathcal{V}$ est le vocabulaire pour lequel il existe une représentation par pl. Ce vocabulaire dépend du corpus d'apprentissage. L'expansion d'une requête Q est donc réalisée par la recherche de termes proches dans le vocabulaire $\mathcal{V}$.

Dans la suite, comme nous fixons a priori tous les paramètres d'apprentissage et que nous nous focalisons sur le corpus d'apprentissage, le paramètre p sera limité au corpus de documents sur lequel les plongements sont appris.

3.1. Expansion locale: terme par terme

Une première approche est celle proposée par (Almasri *et al.*, 2016). Elle se base sur l'utilisation des k plus proches voisins selon une similarité cosinus des vecteurs de termes, qui est équivalente à un produit scalaire des vecteurs normalisés correspondants. L'idée est qu'un terme de l'expansion de requête doit être lié fortement à, au moins, un terme de la requête. De manière plus formelle, nous supposons qu'une requête Q est composée de couples $(q_i, w_{q_i})_{0 \leq i \leq |Q|}$. Dans le cas d'une requête classique, chaque poids w_{q_i} est égal à 1. Dans ce cas, l'expansion $EXP_{loc}(Q, p)$ de la requête initiale Q, avec un plongement pl_p de paramètres p, s'exprime par l'ensemble des paires suivant :

$$EXP_{loc}(Q,p) = \bigcup_{q_i \in Q.termes} \{(t, \alpha_{loc} * Sim(pl_p(t), pl_p(q_i)))| \\ t \in \mathcal{V}, pl_p(t) \in NN_k(pl_p(q_i))\} \quad (2)$$

Les paramètres de l'expression ci-dessus sont les suivants: α_{loc} dénote le poids du terme dans l'expansion (classiquement moins important que les poids des termes initiaux de la requête); k dénote le nombre de plus proches voisins sélectionnés; la similarité $Sim(.,.)$ mesure la proximité entre les vecteurs de deux termes dans l'espace de plongements (une similarité classique (Mikolov *et al.*, 2013) est le $cosinus$ des angles des deux vecteurs). Dans l'expression ci-dessus, la notation $Q.termes$ désigne l'ensemble des mots de la requête $\{q|(w, w_q) \in Q\}$.

3.2. Expansion globale

La seconde approche étudiée repose sur la recherche de termes similaires à toute la requête, elle est similaire à celle proposée par (Almasri, 2017) et proche de cite-RoyPMG16. Dans ce cas, l'expression $EXP_{glob}(Q, p)$ s'exprime de la manière suivante :

$$EXP_{glob}(Q,p) = \left\{ \left(t, \alpha_{glob} * Sim\left(pl_p(t), \sum_{q_i \in Q.termes} pl_p(q_i) \right) \right) \middle| \right. \\ \left. t \in V, pl_p(t) \in NN_k \left(\sum_{q_i \in Q.termes} pl_p(q_i) \right) \right\} \quad (3)$$

Dans l'expression ci-dessus : α_{glob} dénote le poids du terme dans l'expansion. On constate de cette expression qu'un terme ne peut apparaître qu'une fois dans l'expansion. Dans cette configuration, on recherche des termes en relation avec tous les termes de la requête (en fait le vecteur "moyen" de la requête comme on utilise des similarités cosinus).

3.3. Requête étendue

Une fois l'une ou l'autre des expansions effectuées, la requête étendue finale $EXP(Q,p)$ est une fusion de la requête initiale Q et de son expansion $EXP_x(Q,p)$ avec $x = loc$ ou $x = glob$.

Cette fusion a comme objectif d'intégrer de nouveaux termes, mais peut également renforcer un terme existant dans la requête (car il est proche d'autres termes de cette requête). De plus, si un même terme étend plusieurs termes de la requête initiale, alors il va être renforcé (par ajout de chacune des pondérations par terme de la requête).

L'expression finale de $EXP(Q,p)$ est la suivante:

$$EXP(Q,p) = \left\{ \left(t, \sum_{(t,b)\in Q} b + \sum_{(t,c)\in EXP_x(Q,p)} c \right) \middle| \; t \in Q.termes \cup EXP_x(Q,p).termes \right\}$$

Comme nous l'avons dit précédemment, si tous les paramètres d'apprentissage sont fixés, le paramètre p qui va varier dans nos expérimentations se limite au corpus d'apprentissage des plongements.

4. Expérimentations

4.1. Corpus de test

Nous avons testé nos propositions sur le corpus CLEF CMC 2016 sur la recherche de tweets intitulée "Timeline Illustration" (sous-tâche 3, (Goeuriot *et al.*, 2016)). Le corpus complet de CMC contient plus de 50 millions de tweets (contenant le mot "festival" et d'autres termes spécifiques aux festivals considérés). Celui que nous utilisons pour la tâche 3 porte sur deux festival musicaux : "Les vieilles charrues" et les "Transmusicales", il en contient 244 000.

53 requêtes sont évaluées dans cette sous-tâche. Ces requêtes sont des événements d'un jour entier de chaque festival (par exemple un concert en particulier). Par exemple, la requête 1 porte sur le concert de "Khun Narin's Electric" durant les Transmusicales, le 4 décembre 2015 à 16h30 :

```
<topic>
<id>1</id>
<title>Khun Narin's Electric</title>
<festival>Transmusicales</festival>
<begindate>04/12/15-14:00</begindate>
<enddate>04/12/15-16:30</enddate>
</topic>
```

Les pertinences ont été établies manuellement sur le *pool*. Les mesures d'évaluations sont classiques : MAP, précision à 5, 10 et 30 documents. On signale cependant que dans notre cas, les apports les plus notables de notre étude sont obtenus sur les valeurs de précision à 5 documents, nous insisterons donc davantage sur ces valeurs dans la suite.

4.2. Système de recherche d'information

Pour toutes les expérimentations reportées ici, nous avons utilisé le système de recherche d'information Terrier (V 4.0) (Macdonald *et al.*, 2012). Nous nous sommes basés sur le modèle BM25 avec les paramètres par défaut proposés par ce système ($k1 = 1,2; b = 0,75; k3 = 8$). Dans ce cadre, la pondération des termes de la requête est prise en compte durant le calcul de correspondance par le système. Comme notre corpus est composé de documents en français, nous avons par ailleurs utilisé les prétraitements classiques : antidictionnaire français et troncature de Porter [2] pour la langue française, proposés par Terrier.

4.3. Corpus d'entraînement

Nous avons réalisé des expérimentations en utilisant 4 corpus pour générer les plongements de mots :

WF : Wikipedia en français - Il est composé de plus de 1000000 articles, rapatriés du *dump* wikipedia 2017. Les caractéristiques de ce corpus sont les suivantes : il couvre de nombreux sujets, et les pages peuvent également être assez variées en terme de contenu;

WMF : Wikipedia Musical en français - Ce corpus est un sous-ensemble du premier. Il est composé à partir d'une liste d'environ 55000 artistes pour lesquels nous avons extrait de Wikipedia en français les pages, générant un total d'environ 45000 articles Wikipedia. Les articles sont donc spécifiques à la musique, mais les pages peuvent être assez disparates;

2. Code : http://snowball.tartarus.org/algorithms/french/stemmer.html

TF : Tweets généraux en français - 50 Millons de tweets de sous-tâche 2 de la campagne d'évaluation CLEF 2016 CMC workshop. Ces tweets couvrent de nombreux sujets, mais sont courts;

TMF : Tweets musicaux - Les tweets de la tâche 3 de la campagne d'évaluation CLEF 2016. Ce corpus contient des tweets de 2 festivals de musique, pour un total de 244000 tweets.

Le tableau 1 résume les informations sur ces corpus. Nous y indiquons en particulier dans quelle mesure les documents des corpus d'entraînement couvrent les mêmes sujets que ceux du corpus de RI, dans quelle mesure les documents d'un corpus d'apprentissage sont similaires à ceux du corpus de RI, et leurs quantités. Ces quantités sont en nombres de documents.

Tableau 1. Caractéristiques des corpus d'entraînement par rapport au corpus de tweets musicaux du SRI.

Corpus	Adéquation des sujets	Adaptation des types de documents	#docs
WF	−	−	1 M
WMF	+	−	45000
TF	−	+	50 M
TMF	+	+	245000

4.4. Paramètres étudiés

4.4.1. Prétraitement des corpus

Nous utilisons les corpus de deux manières différentes : nous les utilisons d'un côté tels quels (sans prétraitement), et de l'autre en les prétraitant suivant une approche classique de RI par application d'un anti-dictionnaire et de la troncature de Porter (en utilisant le même antidictionnaire et le même code de troncature que le système de recherche d'information). Nous présentons en table 2 les informations liées à la taille des vocabulaires de chacun des quatre corpus, suivant ou non l'utilisation des prétraitements. Nous constatons que l'application de prétraitements pour les pages wikipedia amène à des réductions du nombre de termes de 21% pour WF et 14% pour WMF, et sur les tweets de 10% pour TF et 24% pour TMF. Ceci souligne que les corpus considérés (pages wikipedia et tweets) ne se comportent pas de la même manière en fonction des prétraitements. Dans la suite, nous dénotons ces variantes en indiquant les noms des corpus par ✓ ou ∅, suivant que nous y appliquons des prétraitements ou non.

4.4.2. Apprentissage des plongements

Nous avons utilisé l'outil *word2vec* [3] pour apprendre les plongements. Nous avons utilisé comme architecture le modèle CBOW de ce système. Ce choix a été fait car,

3. https://github.com/dav/word2vec

Tableau 2. Variantes des corpus d'entraînements.

Corpus	Prétraitement	#termes
WF	∅	759 488
WF	✓	596 540
WMF	∅	63 106
WMF	✓	43 786
TF	∅	1 148 947
TF	✓	1 028 488
TMF	∅	21 920
TMF	✓	16 705

d'après (Mikolov *et al.*, 2013), cette version des plongements possède l'avantage d'être rapide. Les autres hyperparamètres de l'apprentissage que nous avons utilisés sont, à défaut d'information additionnelle, ceux par défaut préconisés par ce logiciel présentés dans le tableau 3. Il est à noter que l'utilisation de sous-echantillonnage par word2vec, qui vise à ôter des instances de termes très courants, joue un rôle assez similaire à l'anti-dictionnaire des prétraitements ci-dessus. Cependant, les prétraitements éliminent toutes les occurrences des termes de l'antidictionnaire, ce que ne fait pas le sous-échantillonnage.

Tableau 3. Valeurs des hyperparamètres utilisés par word2vec.

Hyperparamètre	Valeur
Dimension de l'espace de plongement	200
Fenêtre	8
Nombre d'itérations	15

4.4.3. Paramètres d'expansion

Les paramètres d'expansion que nous avons utilisés sont :

– les deux variantes du calcul des termes d'expansion : expansion globale et expansion locale comme décrits plus haut;

– le nombre de termes d'expansion, dans l'ensemble [1, 5]. Ce paramètre correspond au k dans les parties précédentes;

– le poids assigné aux termes étendus. Comme nous l'avons dit précédemment, nous utilisons dans cet article une pondération fixe pour chaque terme étendu. Cette pondération, de manière classique en recherche d'information (cf. Rocchio par exemple), est habituellement moins grande que celle des termes initiaux de la requête, car ces extensions sont calculées et donc moins fiables. Nous avons étudié ici 3 pondérations [0,1; 0,2; 0,3] pour chacun des paramètres α_{loc} et α_{glob}.

4.5. Résultats

Nous avons particulièrement étudié ici quatre éléments prépondérants :

– L'impact des prétraitements et des ensembles d'apprentissage, et plus particulièrement l'application d'outils classique de RI : anti-dictionnaire et troncature de mots;

– L'impact des paramètres de pondération des expansions, afin de déterminer l'importance à donner aux termes ajoutés à la requête;

– L'impact de la taille des expansions, afin de déterminer le nombre de termes qui donnent les meilleurs résultats.

Nous donnons tout d'abord les résultats obtenus sans expansion de requête dans le tableau 4.

Tableau 4. Les résultats sans expansion de requêtes.

MAP	Recip_rank	p@5	p@10	p@30
0.0062	$0,4499$	$0,2718$	$0,2385$	$0,2094$

Nous nous concentrons dans la suite sur les résultats de précision à 5 documents (p@5), ce qui permet de mesurer les différences de qualité sur les premiers résultats fournis en réponse.

4.5.1. Impact des prétraitements et des corpus d'apprentissage

Nous décrivons dans les tableaux 5 et 6 l'impact des prétraitements du corpus d'entraînement (i.e., anti-dictionnaire + troncature) sur les expansions des requêtes. Pour cela, nous nous concentrons sur les paramètres $\alpha_{loc} = 0,3$ et $\alpha_{glob} = 0,3$ (cf. partie 4.5.2), et nous regardons les résultats en faisant varier le paramètre k de 0 (i.e., sans expansion) à 5, tous les autres paramètres étant fixés. Dans ces tableaux, nous indiquons les meilleurs résultats par nombre de termes d'expansion (i.e., par colonne) en gras, et nous indiquons en souligné les meilleurs résultats par corpus considéré (i.e., par ligne).

Nous constatons que, quelle que soit l'expansion considérée, les résultats en termes de P@5 sont meilleurs que ceux sans expansion de requête (cf tableau 4). De manière globale, les résultats par expansion localedu tableau 5 donnent des résultats sensiblement supérieurs à ceux avec expansion globale du tableau 6. Ceci peut provenir du fait que les termes des requêtes ne sont pas forcément liés dans les plongements, ce qui fait qu'utiliser de manière conjointe tous les termes n'est pas la meilleure méthode.

On se rend compte également que l'expansion avec 5 termes donne des résultats au moins égaux à l'expansion avec un seul terme, quelle que soit la configuration considérée, à l'exception de l'expansion globale pour $WF_{\emptyset}$. Ceci veut dire qu'il ne

3. † and ∘ dénotent une différence statistiquement significative en P@5 avec le résultat sans expansion de requête, et la meilleure de configuration testée sur la même ligne (souligné). Le test de Student pairé bilatéral avec seuil de significativité de 0.1 est utilisé.

Tableau 5. Les prétraitements par corpus d'apprentissage, avec $k \in [1, 5]$ pour EXP_{loc} [2]

Corpus	p@5 | k=1	p@5 | k=2	p@5 | k=3	p@5 | k=4	p@5 | k=5
$WF_{\emptyset}$	$\underline{0,2895}$	$\underline{0,2895}$	$\underline{0,2895}$	$\underline{0,2895}$	$\underline{0,2895}$
$WF_{\checkmark}$	$0,2842$	$\underline{0,2947}$	$\underline{0,2947}$	$\underline{0,2947}$	$\underline{0,2947}$
$WMF_{\emptyset}$	$\underline{0,2789}$	$\underline{0,2789}$	$\underline{0,2789}$	$0,2737$	$0,2737$
$WMF_{\checkmark}$	$0,2789$	$0,2789$	$\underline{0,2842}$	$\underline{0,2842}$	$\underline{0,2842}$
$TF_{\emptyset}$	$\underline{0,2737}$	$\underline{0,2737}$	$0,2737$	$0,2737$	$0,2737$
$TF_{\checkmark}$	$0,2789\circ$	$0,2895\circ$	$0,3105$	$0,3211$	$\underline{0,3263}$
$TMF_{\emptyset}$	$\mathbf{0,2947}\circ$	$\mathbf{0,3158}$	$\mathbf{0,3158}$	$\mathbf{0,3316}$	$\mathbf{0,3421}\dagger$
$TMF_{\checkmark}$	$0,2895$	0.2947	$0,3105$	$0,3105$	$\underline{0,3211}$

Tableau 6. Les prétraitements par corpus d'apprentissage, avec $k \in [1, 5]$ pour EXP_{glob} [2].

Corpus	p@5 | k=1	p@5 | k=2	p@5 | k=3	p@5 | k=4	p@5 | k=5
$WF_{\emptyset}$	$\underline{0,2842}$	$\underline{0,2842}$	$\underline{0,2842}$	$\underline{0,2842}$	$0,2769$
$WF_{\checkmark}$	$0,2789$	$\underline{0,2842}$	$\underline{0,2842}$	$\underline{0,2842}$	$\underline{0,2842}$
$WMF_{\emptyset}$	$0,2737$	$0,2737$	$0,2737$	$0,2737$	$\underline{0,2789}$
$WMF_{\checkmark}$	$\underline{0,2737}$	$\underline{0,2737}$	$\underline{0,2737}$	$\underline{0,2737}$	$\underline{0,2737}$
$TF_{\emptyset}$	$\underline{0,2737}$	$\underline{0,2737}$	$\underline{0,2737}$	$0,2737$	$0,2737$
$TF_{\checkmark}$	$0,2737$	$0,2789$	$\underline{0,3000}$	$\underline{0,3000}$	$0,2974$
$TMF_{\emptyset}$	$\mathbf{0,2895}\circ$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3158}\dagger$
$TMF_{\checkmark}$	$0,2842$	$0,2842$	$0,2947$	$0,2895$	$\underline{0,3053}$

faut pas se limiter à une expansion très courte, mais aussi que pour les plongements appris sur corpus d'apprentissage très généraux il faut limiter ce nombre.

Pour les deux types d'expansion, on constate que l'utilisation du corpus d'apprentissage qui correspond au corpus de recherche donne les meilleurs résultats, et ceci malgré le fait que ce corpus soit relativement petit.

Avec les expansions locales, on constate que l'utilisation des prétraitements sur les corpus WF, WMF et TF donne des résultats aussi bons ou meilleurs que les corpus non-prétraités. Les prétraitements que l'on a utilisés ne sont donc pas compatibles avec l'apprentissage des plongements, ce qui peut s'expliquer par le fait qu'une même troncature peut apparaître dans de nombreux contextes qui ne sont pas discriminés.

Dans les deux cadres d'expansions utilisés, les calculs de significativités statistiques obtenus permettent d'obtenir des différences significatives avec l'approche non-étendue uniquement pour le corpus de tweets musicaux non-prétraités $TMF_{\emptyset}$ pour une expansion de $k = 5$. Ceci renforce encore le fait qu'apprendre le -s plongements sur le corpus de recherche non-prétraité est utile.

4.5.2. Impact des pondérations α_{loc} et α_{glob}

Nous étudions ici les résultats obtenus en fonction de la pondération préfixée pour les expansions de requêtes. Pour cela, nous nous concentrons sur les meilleures configurations pour chaque type d'expansion (locale ou globale), c'est-à-dire $TMF_{\emptyset}$, sans prétraitement. Les résultats sont présentés dans les tableaux 7 et 8, en faisant varier α dans la liste de valeurs $\{0,1; 0,9\}$.

Tableau 7. L'impact des valeurs $\alpha_{loc} \in [0,1,0,9]$ pour EXP_{loc} avec apprentissage $TMF_{\emptyset}$, pour $k \in [1,5]$.

α_{loc}	p@5 | k=1	p@5 | k=2	p@5 | k=3	p@5 | k=4	p@5 | k=5
$0,1$	$0,2789$	$0,3000$	$0,3053$	$0,3053$	$0,3158$
$0,2$	$0,2842$	$0,3053$	$0,3105$	$0,3158$	$0,3263$
$0,3$	$0,2947$	$0,3158$	$0,3158$	$\mathbf{0,3316}$	$\mathbf{0,3421}$
$0,4$	$0,3105$	$\mathbf{0,3316}$	$\mathbf{0,3316}$	$\mathbf{0,3316}$	$\mathbf{0,3421}$
$0,5$	$\mathbf{0,3158}$	$\mathbf{0,3316}$	$\mathbf{0,3316}$	$\mathbf{0,3316}$	$\mathbf{0,3421}$
$0,6$	$\mathbf{0,3158}$	$\mathbf{0,3316}$	$\mathbf{0,3316}$	$\mathbf{0,3316}$	$\mathbf{0,3421}$
$0,7$	$\mathbf{0,3158}$	$0,3263$	$0,3263$	$\mathbf{0,3316}$	$\mathbf{0,3421}$
$0,8$	$\mathbf{0,3158}$	$0,3263$	$0,3263$	$\mathbf{0,3316}$	$\mathbf{0,3421}$
$0,9$	$\mathbf{0,3158}$	$0,3263$	$0,3263$	$\mathbf{0,3316}$	$\mathbf{0,3421}$

Tableau 8. L'impact des valeurs α_{glob} pour EXP_{glob} avec apprentissage $TMF_{\emptyset}$, pour $k \in [1,5]$.

α_{glob}	p@5 | k=1	p@5 | k=2	p@5 | k=3	p@5 | k=4	p@5 | k=5
$0,1$	$0,2789$	$0,3000$	$0,3053$	$0,3053$	$0,3105$
$0,2$	$0,2842$	$0,3053$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3158}$
$0,3$	$0,2895$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3158}$
$0,4$	$0,2895$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3158}$
$0,5$	$\mathbf{0,2947}$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3158}$
$0,6$	$\mathbf{0,2947}$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3158}$
$0,7$	$\mathbf{0,2947}$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3158}$
$0,8$	$\mathbf{0,2947}$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3158}$
$0,9$	$\mathbf{0,2947}$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3105}$	$\mathbf{0,3158}$

On constate de ces deux tableaux que, d'une part les expansions avec le $\alpha = 0,1$ donnent la même qualité de réponses pour les deux approches, et d'autre par de manière générale les résultats atteignent un maximum, aussi bien pour α_{loc} ou α_{glob}, entre 0,2 et 0,4 , puis forment un plateau à cette valeur. Le comportement de l'expansion globale, dans le tableau 8, ce plateau arrive pour des valeurs de α_{glob} plus faibles, et la différence de résultats entre les valeurs de α_{glob} égales à 0,2 et 0,3 est moins notable. Dans ce cas, la valeur d'importance des termes ajoutés est donc moins cruciale. Ce constat est moins présent pour les expansions locales. On remarque cependant que pour les sélections des 2 ou 3 meilleurs termes dans l'expansion locale, le fait d'accorder une plus grande importance aux expansions dégrade la qualité des

réponses (troisièmes et quatrièmes colonnes du tableau 7. Un élément assez remarquable est visible entre la deuxième colonne de 8 pour les $\alpha_{loc} \geq 0,4$ et la dernière de 8 : les valeurs de p@5 obtenues sont très corrélées. On peut mettre ici en avant le fait que l'expansion locale considérée ajoute un terme par terme de la requête, alors que l'expansion globale considérée ajoute 5 termes. Dans le cas de requêtes avec plusieurs termes, l'expansion est alors d'une taille proche. Une étude sur le nombre de termes d'expansion serait à faire dans les travaux futurs.

Dans les cas listé ci-dessus, on retrouve également le fait que plus les expansions sont grandes (dans la plage fixée entre 1 et 5), meilleur est le résultat.

4.6. Discussion

Un élément important que nous signalons est que, comme le corpus de tweets musicaux est relativement petit, on pourrait supposer que l'apprentissage des plongements ne serait pas intéressant. Les résultats que nous obtenons dans le cadre de nos expérimentations montrent le contraire. Il semble donc préférable d'utiliser un corpus petit mais bien adapté, plutôt qu'un corpus plus grand mais moins spécifique. Comme nous nous situons dans le cas spécifique de la prise en compte des plongements pour la recherche d'information "classique", nous devons aussi garder à l'esprit que nous nous reposons toujours sur l'intersection entre les termes de la requête et les termes des documents. Il en résulte que les termes de l'expansion seront d'autant plus intéressants qu'ils indexent les documents. Utiliser le corpus de recherche pour l'apprentissage des plongements possède donc l'avantage indéniable d'utiliser des termes du vocabulaire d'indexation pour les expansions, ce qui n'est pas garanti sinon, surtout dans le cas de corpus de textes spécifiques.

Comme l'a montré (Billerbeck, Zobel, 2004), l'expansion de requête n'est pas efficace dans tous les contextes, ni sur toutes les requêtes. Une adaptation des paramètres d'expansion aux requêtes et tâches pourrait permettre de limiter le bruit et d'améliorer la recherche de microblogs pertinents.

5. Conclusion

Dans cet article, nous avons étudié l'influence du corpus d'apprentissage utilisé pour un plongement de mots sur la recherche de microblogs culturels. Pour cela, nous avons utilisé deux approches relativement simples pour réaliser des expansions de requêtes à partir de ces plongements pour étudier leur impact. L'avantage de cette simplicité était d'éviter d'avoir à faire varier plusieurs paramètres à la fois lors des expérimentations. Les résultats obtenus dans le cadre de la recherche de tweets culturels montrent qu'il est préférable d'apprendre les plongements sur le corpus utilisé par le système de recherche d'information, même si celui-ci est petit par rapport aux travaux de l'état de l'art. Une explication possible est que, comme les plongements sont appris sur le même corpus que celui du SRI, il y a moins de de problèmes de mots hors vocabulaire pour la prise en compte des expansions.

Dès lors, une question qui se pose est de pouvoir profiter conjointement de corpus vastes d'un côté, et de corpus spécifiques de l'autre. Une étude future portera sur la comparaison entre : a) réaliser un premier apprentissage sur un corpus géréral comme Wikipedia, puis continuer l'apprentissage sur un second, ou bien b) uniquement concaténer les corpus en un seul puis réaliser un seul apprentissage sur ce corpus final. Une autre direction pour utiliser plusieurs corpus d'apprentissage serait de s'inspirer de fusion tardive, et d'étudier comment se comportent des expansions intégrant soit les expansions de plusieurs plongements provenant de plusieurs corpus d'apprentissage soit en fusionnant les vecteurs de plongements appris de plusieurs corpus comme dans (Ghannay *et al.*, 2016). Ces études feront l'objet de travaux futurs.

Remerciements

Ce travail a été partiellement financé par les projets : LIG Emergence 2017 *Tonel* et LIG Emergence 2018 *Arosoir*.

Bibliographie

Almasri M. (2017). *Réduire la probabilité de disparité des termes en exploitant leurs relations sémantiques*. Thèse de doctorat non publiée, Université Grenoble Alpes.

Almasri M., Berrut C., Chevallet J. (2016). A comparison of deep learning based query expansion with pseudo-relevance feedback and mutual information. In *Advances in information retrieval - 38th european conference on IR research, ECIR 2016, padua, italy, march 20-23, 2016. proceedings*, p. 709–715.

Bengio Y., Ducharme R., Vincent P., Jauvin C. (2003). A neural probabilistic language model. , vol. 3, p. 1137–1155. Consulté sur http://www.iro.umontreal.ca/~lisa/pointeurs/BengioDucharmeVincentJauvin_jmlr.pdf

Billerbeck B., Zobel J. (2004). Questioning query expansion: an examination of behaviour and parameters. In *Adc '04: Proceedings of the 15th australasian database conference*, p. 69–76. Darlinghurst, Australia, Australia, Australian Computer Society, Inc.

Chiu B., Korhonen A., Pyysalo S. (2016). Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st workshop on evaluating vector-space representations for nlp, repeval@acl 2016, berlin, germany, august 2016*, p. 1–6. Consulté sur https://doi.org/10.18653/v1/W16-2501

Diaz F., Mitra B., Craswell N. (2016, August). Query expansion with locally-trained word embeddings. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, p. 367–377. Berlin, Germany, Association for Computational Linguistics. Consulté sur http://www.aclweb.org/anthology/P16-1035

Ghannay S., Favre B., Estève Y., Camelin N. (2016). Word embedding evaluation and combination. In *10th edition of the Language Resources and Evaluation Conference (LREC 2016)*. Portorož, Slovenia. Consulté sur https://hal.archives-ouvertes.fr/hal-01433185

Gimpel K., Schneider N., O'Connor B., Das D., Mills D., Eisenstein J. *et al.* (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of*

the 49th annual meeting of the association for computational linguistics: Human language technologies: Short papers - volume 2, p. 42–47. Stroudsburg, PA, USA, Association for Computational Linguistics. Consulté sur http://dl.acm.org/citation.cfm?id=2002736.2002747

Goeuriot L., Mothe J., Mulhem P., Murtagh F., SanJuan E. (2016). Overview of the CLEF 2016 cultural micro-blog contextualization workshop. In *Experimental IR meets multilinguality, multimodality, and interaction - 7th international conference of the CLEF association, CLEF 2016, évora, portugal, september 5-8, 2016, proceedings*, p. 371–378. Consulté sur https://doi.org/10.1007/978-3-319-44564-9_30

Iacobacci I., Pilehvar M. T., Navigli R. (2015). Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing of the asian federation of natural language processing, ACL 2015, july 26-31, 2015, beijing, china, volume 1: Long papers*, p. 95–105. Consulté sur http://aclweb.org/anthology/P/P15/P15-1010.pdf

Jabeur L. B., Tamine L., Boughanem M. (2012). Intrégration des facteurs temps et autorité sociale dans un modèle bayésien de recherche de tweets. In *CORIA (conférence en recherche d'infomations et applications) - CORIA 2012, 9th french information retrieval conference, bordeaux, france, march 21-23, 2012. proceedings*, p. 301–316. Consulté sur https://doi.org/10.24348/coria.2012.301

Kuzi S., Shtok A., Kurland O. (2016). Query expansion using word embeddings. In *Proceedings of the 25th acm international on conference on information and knowledge management*, p. 1929–1932. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/2983323.2983876

Macdonald C., McCreadie R., Santos R. L., Ounis I. (2012). From puppy to maturity: Experiences in developing terrier. *Proc. of OSIR at SIGIR*, p. 60–63.

Mikolov T., Chen K., Corrado G., Dean J. (2013). Efficient estimation of word representations in vector space. *CoRR*, vol. abs/1301.3781. Consulté sur http://arxiv.org/abs/1301.3781

Mitra B., Craswell N. (2017). Neural models for information retrieval. *CoRR*, vol. abs/1705.01509. Consulté sur http://arxiv.org/abs/1705.01509

Mohammad S. M., Sobhani P., Kiritchenko S. (2017, juin). Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, vol. 17, n° 3, p. 26:1–26:23. Consulté sur http://doi.acm.org/10.1145/3003433

Roy D., Paul D., Mitra M., Garain U. (2016). Using word embeddings for automatic query expansion. In *Neu-ir '16 sigir workshop on neural information retrieval.* Consulté sur http://arxiv.org/abs/1606.07608

Twitter. (2018). *Tweet data dictionaries.* https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/entities-object.

Zhang Y., Rahman M. M., Braylan A., Dang B., Chang H., Kim H. *et al.* (2016). Neural information retrieval: A literature review. *CoRR*, vol. abs/1611.06792. Consulté sur http://arxiv.org/abs/1611.06792