Contextualisation automatique de Tweets à partir de Wikipédia

Romain Deveaud [1] — **Florian Boudin** [2]

[1] *LIA - Université d'Avignon*
romain.deveaud@univ-avignon.fr
[2] *LINA - Université de Nantes*
florian.boudin@univ-nantes.fr

RÉSUMÉ. *Les réseaux sociaux sont au centre des communications sur internet et une grande partie des échanges communautaires se fait à travers eux. Parmi eux, l'apparition de Twitter a donné lieu à la création d'un nouveau type de partage d'informations où les messages sont limités à 140 caractères. Les utilisateurs de ce réseau s'expriment donc succinctement, souvent en temps réel à partir d'un smartphone, et la teneur des messages peut parfois être difficile à comprendre sans contexte. Nous proposons dans cet article une méthode permettant de contextualiser automatiquement des Tweets en utilisant des informations provenant directement de l'encyclopédie en ligne Wikipédia, avec comme but final de répondre à la question : « De quoi parle ce Tweet ? ». Nous traitons ce problème comme une approche de résumé automatique où le texte à résumer est composé d'articles Wikipédia liés aux différentes informations exprimées dans un Tweet. Nous explorons l'influence de différentes méthodes de recherche d'articles liés aux Tweets, ainsi que de plusieurs caractéristiques utiles pour la sélection des phrases formant le contexte. Nous évaluons notre approche en utilisant la collection de la tâche Tweet Contextualization d'INEX 2012 et donnons un aperçu sur ce qui caractérise une phrase importante pour déterminer le contexte d'un Tweet.*

ABSTRACT. *Social networks are central in nowadays internet communication and community exchanges. The emergence of Twitter led to the creation of a new tool for sharing information, where messages are bound to 140 characters. Publications on this social network are short and straightforward and often sent in real time from mobile phones, which make it difficult to apprehend without some kind of context. We propose in this paper a method allowing to automatically contextualize Tweets by using information coming from Wikipedia. We treat this problem as an automatic summarization task, where the text to resume is composed of Wikipedia article that discuss the various pieces of information appearing in a Tweet. We explore the influence of various Tweet-related articles retrieval methods as well as several features for sentence extraction. We evaluate our approach using the test collection from the INEX 2012 Tweet Contextualization track and provide some insights on what makes a contextually important sentence.*

MOTS-CLÉS : *Contexte informatif, résumé par extraction, Twitter, Wikipédia*

KEYWORDS: *Topical context, sentence extraction, Twitter, Wikipedia*

1. Introduction

La grande démocratisation de l'accès à internet et l'avènement des smartphones ont changé le paysage virtuel et la nature des échanges entre les personnes. L'information n'attend plus forcément d'être trouvée par quelqu'un ayant un besoin précis, elle vient directement à nous. Au centre de ce phénomène, les réseaux sociaux sont un média privilégié pour la diffusion de contenu à grande échelle (Bakshy *et al.,* 2012). Les utilisateurs sont reliés par des connections de natures diverses (professionnelles, personnelles, publicitaires...) et s'échangent des informations en temps réel sur le monde qui les entoure. Twitter fait partie de ces réseaux sociaux et favorise des échanges de messages très courts. Quand il se connecte à Twitter, l'utilisateur doit répondre à la question « Quoi de neuf ? ». La réponse à cette question doit faire moins de 140 caractères et est appelée un *Tweet*. De par sa taille, un Tweet est naturellement ambigu et souvent sous-spécifié, ce qui peut rendre la compréhension compliquée pour une personne ne possédant pas le contexte approprié. Ce contexte peut être formé de phrases récupérées sur le Web (ou toute autre source) et réunies afin d'éclairer les lecteurs d'un Tweet sur sa nature et sur les concepts informatifs mis en jeu.

Nous plaçons notre étude dans le cadre d'un scénario mobile où un utilisateur va lire des Tweets (ou autres messages courts) sur son smartphone. Le contexte d'un Tweet doit donc être court afin de pouvoir être affiché de façon pratique sur un écran de téléphone. La tâche *Tweet Contextualization* d'INEX[1] propose un cadre expérimental permettant d'évaluer la contextualisation de Tweets réalisée à l'aide de phrases issues de Wikipédia. La collection de test est composée d'un ensemble statique d'articles Wikipédia, de Tweets et de phrases contextuelles de référence sélectionnées par les organisateurs.

Notre approche de la contextualisation met en jeu successivement des techniques de Recherche d'Information (RI) et de résumé automatique. Tout d'abord, nous cherchons à améliorer la compréhension du Tweet en récupérant des articles Wikipédia liés à celui-ci. Ces derniers sont susceptibles de contenir des passages informatifs pour la construction du contexte du Tweet. Ensuite, nous considérons la formation du contexte comme une tâche de résumé automatique multi-documents, où il s'agit de résumer les articles Wikipédia retournés. Nous présentons dans cet article le modèle de RI puis l'approche de résumé automatique qui constituent notre système de contextualisation, puis nous évaluons notre approche en utilisant l'ensemble de données issu de la tâche *Tweet Contextualization* d'INEX 2012 (SanJuan *et al.,* 2012).

2. Travaux précédents

Le problème de contextualisation de messages courts est émergent et se situe aux confluents de la Recherche d'Information ciblée et du résumé automatique. La tâche *Tweet Contextualization* de la campagne d'évaluation INEX 2012 est la première à

1. `https://inex.mmci.uni-saarland.de/`

proposer un cadre d'évaluation formel pour ce type de problématique et a été suivie par de nombreux participants. Allant dans le même sens, une nouvelle tâche de *Temporal Summarization* va faire son apparition à TREC pour l'année 2013. Le but sera ici de produire des résumés évoquant des grands événements (ouragans, élections...) et d'ordonner les différentes phrases chronologiquement.

Au cours de la dernière décennie, de nombreux chercheurs se sont penchés sur la problématique du résumé automatique. La quasi-totalité des approches proposées recourent à des méthodes d'extraction où il s'agit d'identifier les unités textuelles, le plus souvent des phrases, les plus importantes des documents. Les phrases les plus pertinentes sont ensuite assemblées pour générer le résumé.

De nombreuses méthodes ont été utilisées pour évaluer l'importance des phrases, e.g. (Barzilay *et al.,* 1997, Radev *et al.,* 2004). Parmi elles, les méthodes basées sur les modèles de graphes (Mihalcea, 2004) donnent de bons résultats. L'idée est de représenter le texte sous la forme d'un graphe d'unités textuelles (phrases) inter-connectées par des relations de similarité. Des algorithmes d'ordonnancement tels que PAGERANK (Page *et al.,* 1999) sont ensuite utilisés pour sélectionner les phrases les plus centrales dans le graphe.

Le résumé automatique orienté (Dang, 2005) est probablement la tâche qui se rapproche le plus de la contextualisation automatique. Il s'agit de générer un résumé répondant à un besoin utilisateur exprimé sous la forme d'une requête. Une grande partie des approches proposées reposent sur des méthodes de résumé automatique existantes et y ajoutent divers critères de pertinence par rapport à la requête, e.g. (Boudin *et al.,* 2008). Parmi les différentes méthodes utilisées pour estimer la pertinence des phrases, plusieurs modèles issus de la RI donnent de bons résultats (Wei *et al.,* 2008).

3. Recherche de phrases candidates contextuelles issues de Wikipédia

Dans le cadre de la tâche Tweet Contextualization d'INEX, le contexte d'un Tweet est défini par un texte composé de 500 mots au maximum et dont les phrases sont issues de Wikipédia. La Figure 1 illustre la méthodologie que nous utilisons. Cette section détaille le processus de sélection des phrases candidates qui pourront composer le contexte.

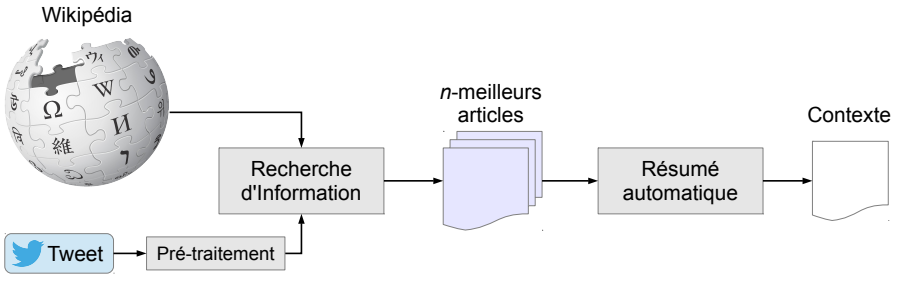

Figure 1. *Méthodologie de contextualisation d'un Tweet à partir de Wikipédia.*

3.1. *Interprétation des #HashTags et formatage des Tweets*

La première étape que nous effectuons consiste à appliquer un ensemble de prétraitements aux Tweets. Il s'agit de formater le contenu de ces derniers en vue de l'étape de recherche d'information.

Le symbole #, appelé *hashtag*, est utilisé pour signaler des mots-clés ou des sujets dans un Tweet. Il a été créé par les utilisateurs de Twitter comme un moyen permettant de catégoriser leurs messages. Les utilisateurs emploient les *hashtags* avant un mot-clé ou une phrase pertinente (sans espace) de leurs Tweets. Ils agissent comme un moyen de catégorisation et d'étiquetage et sont ainsi des marqueurs d'informations importantes directement fournis par l'auteur. Il semble donc logique de privilégier leur utilisation dans le cadre d'une récupération d'articles Wikipédia liés à un Tweet.

La principale difficulté avec l'utilisation des *hashtags* vient du fait qu'ils sont pour la plupart composés de plusieurs mots concaténés. La figure 2 illustre ce problème avec le hashtag #WhitneyHouston. Dans ce cas précis, il ne serait pas possible pour un système de recherche d'information classique de renvoyer des documents liés à Whitney Houston étant donné que ces deux mots n'apparaissent pas dans le Tweet.

Figure 2. *Exemple d'un Tweet issu de la collection INEX Tweet Contextualization pour l'année 2012.*

Pour résoudre ce problème, nous avons utilisé un algorithme de segmentation automatique de mots basé sur celui présenté dans le chapitre « Natural Language Corpus Data » du livre « Beautiful Data » (Segaran *et al.,* 2009). Nous calculons le découpage le plus probable d'un *hashtag* à l'aide des probabilités d'apparition d'unigrammes et de bigrammes au sein du corpus Bing N-Gram[2]. Ainsi, chaque *hashtag* présent dans le Tweet initial est remplacé par sa version découpée.

Twitter étant un réseau social, l'interaction entre les utilisateurs est au centre de son fonctionnement. Ainsi, un Tweet peut contenir différentes mentions destinées à d'autres personnes, comme par exemple une réponse ou un retweet. Un Tweet réponse commence par un @ suivi du pseudonyme d'un (ou plusieurs) utilisateur(s). Cela

2. `http://web-ngram.research.microsoft.com/info/`

permet notamment de créer une discussion spontanée entre plusieurs personnes. Quant au retweet, il consiste à reposter le Tweet d'une autre personne. Parfois les utilisateurs tapent *RT* au début d'un Tweet pour indiquer qu'ils repostent le contenu d'un autre utilisateur. Ce n'est pas une commande ou une fonction officielle de Twitter, mais cela signifie qu'ils citent le Tweet d'un autre utilisateur. Néanmoins ces différentes mentions n'apportent rien au contenu informatif du Tweet, nous les supprimons donc simplement. Les mots outils sont également supprimés en utilisant la liste standard INQUERY fournie avec le système de recherche d'information Indri[3]. La sortie finale de cette étape de formatage est un Tweet nettoyé, sans mots-outils, ni *hashtags* collés, ni mentions inutiles.

3.2. *Recherche d'articles Wikipédia*

La sélection d'articles Wikipédia apportant des informations contextuelles par rapport à un Tweet est une étape cruciale pour trouver les phrases qui vont former le contexte. Nous présentons dans cette section les différentes méthodes de recherche documentaire que nous utilisons dans nos expériences.

3.2.1. *Modèle de base*

L'une des approches standard de la recherche d'information par modèle de langue se fait avec un modèle de vraisemblance de la requête. Ce modèle mesure la probabilité que la requête puisse être générée à partir d'un document donné, ainsi les documents sont ordonnés en se basant sur cette probabilité. Soit θ_D le modèle de langue estimé en se basant sur un document D, le score d'appariement entre D et une requête $\mathcal{T}$ est défini par la probabilité conditionnelle suivante :

$$P(\mathcal{T}|\theta_D) = \prod_{t \in \mathcal{T}} f_T(t, D) \quad [1]$$

Un des points importants dans le paramétrage des approches par modèle de langue est le lissage des probabilités nulles. Dans ce travail, θ_D est lissé en utilisant le lissage de Dirichlet (Zhai *et al.,* 2004), on a donc :

$$f_T(t, D) = \prod_{t \in \mathcal{T}} \frac{c(t, D) + \mu \cdot P(t|\mathcal{C})}{|D| + \mu}$$

où $c(t, D)$ est le nombre d'occurrences du mot t dans le document D. $\mathcal{C}$ représente la collection de documents et μ est le paramètre du lissage de Dirichlet (nous fixons $\mu = 2500$ tout au long de cet article).

Une des limitations évidente de l'approche par unigramme est qu'elle ne tient pas compte des dépendances ou des relations qu'il peut y avoir entre deux termes adjacents

3. `http://www.lemurproject.org/`

dans la requête. Le modèle MRF (Markov Random Field) (Metzler *et al.,* 2005) est une généralisation de l'approche par modèle de langue et résoud spécifiquement ce problème. L'intuition derrière ce modèle est que des mots adjacents de la requête sont susceptibles de se retrouver proches dans les documents pertinents. Trois différents types de dépendances sont considérés :

1) l'indépendance des termes de la requête (ce qui revient à un modèle de langue standard prenant en compte uniquement les unigrammes),

2) l'apparition exacte de bigrammes de la requête,

3) et l'apparition de bigrammes de la requête dans un ordre non défini au sein d'une fenêtre de mots.

Le modèle propose deux fonctions supplémentaires pour deux autres types de dépendances qui agissent sur les bigrammes de la requête :

$$f_O(t_i, t_{i+1}, D) = \frac{c(\#1(t_i, t_{i+1}), D) + \mu \cdot \frac{c(\#1(t_i, t_{i+1}), \mathcal{C})}{|\mathcal{C}|}}{|D| + \mu}$$

$$f_U(t_i, t_{i+1}, D) = \frac{c(\#uw8(q_i, q_{i+1}), D) + \mu \cdot \frac{c(\#uw8(q_i, q_{i+1}), \mathcal{C})}{|\mathcal{C}|}}{|D| + \mu}$$

La fonction $f_O(q_i, q_{i+1}, D)$ considère la correspondance exacte de deux mots adjacents de la requête. Elle est dénotée par l'indice O. La seconde est dénotée par l'indice U et considère la correspondance non ordonnée de deux mots au sein d'une fenêtre de 8 unités lexicales. Ici, $c(\#1(t_i, t_{i+1}), D)$ est le nombre d'occurrences du bigramme (t_i, t_{i+1}) dans le document D. Comparativement, $c(\#uw8(t_i, t_{i+1}), D)$ est le nombre d'occurences des deux mots de la requête t_i et t_{i+1} au sein d'une fenêtre non ordonnée composée de 8 termes du document D.

Finalement, le score d'un article Wikipédia D par rapport à un Tweet formaté $\mathcal{T}$ est donné par la fonction suivante :

$$\begin{aligned} s_{MRF}(\mathcal{T}, D) \quad = \quad & \lambda_T \prod_{t \in \mathcal{T}} f_T(t, D) \; + \\ & \lambda_O \prod_{i=1}^{|Q|-1} f_O(t_i, t_{i+1}, D) \; + \\ & \lambda_U \prod_{i=1}^{|Q|-1} f_U(t_i, t_{i+1}, D) \end{aligned} \qquad [2]$$

où λ_T, λ_O et λ_U sont des paramètres libres dont la somme est égale à 1. Dans nos expériences nous fixons ces paramètres en suivant les recommandations des auteurs ($\lambda_T = 0,85$, $\lambda_O = 0,10$ et $\lambda_U = 0,05$).

3.2.2. *Intégration de* hashtags

Les *hashtags* peuvent être considérés comme des étiquettes définies manuellement par les auteurs des Tweets. Ce sont par conséquent des marqueurs évidents d'informations importantes. Ils peuvent également être considérés comme des requêtes courtes, sorte d'abréviation du Tweet. Considérons le Tweet $\mathcal{T}$ suivant :

> « All #Airbus #A380 Jumbo Jets Ordered To Be Inspected For Wing Cracks - Neon Tommy : http ://t.co/SofXXzCN »

Le sujet principal est correctement représenté par un ensemble de *hashtags* $H_{\mathcal{T}} = \{"airbus", "a380"\}$. Nous pouvons ainsi le considérer comme une simplification du Tweet ou encore une expression des informations les plus importantes. Un parallèle peut également être fait avec les *topics* de TREC qui sont traditionnellement composés d'une requête courte (2 à 5 mots-clés) et d'une description plus détaillée du besoin d'information (pouvant comprendre plusieurs phrases).

Nous introduisons donc les *hashtags* de façon explicite dans la fonction de score des articles Wikipédia de notre système. Soient un Tweet $\mathcal{T}$ et ses *hashtags* $H_{\mathcal{T}}$, le score d'un article Wikipédia D est donné par :

$$s(\mathcal{T}, H_{\mathcal{T}}, D) = \alpha\, s_{MRF}(H_{\mathcal{T}}, D) + (1-\alpha)\, s_{MRF}(\mathcal{T}, D) \qquad [3]$$

Le paramètre α permet de maintenir la balance entre l'influence des *hashtags* seuls et le Tweet entier. Nous nous plaçons dans le cadre d'une contextualisation en temps réel, et la nature très hétérogène des Tweets ne nous semble pas adaptée pour effectuer un apprentissage *a priori* de ce paramètre. De plus, les *hashtags* peuvent avoir une utilité parfois très limitée voire nulle, comme dans l'exemple suivant :

> « U Just Heard “Hard To Believe" by @andydavis on the @mtv Teen Mom 2 Finale go 2 http ://t.co/iwb2JuL8 for info #iheardítonMTV »

Dans ce cas-ci, « I heard it on MTV » est une phrase d'accroche de type publicitaire et n'apporte rien pour la compréhension du Tweet. L'importance des *hashtags* est donc elle aussi contextuelle et dépend de leur pouvoir discriminant. Nous choisissons d'estimer ce pouvoir discriminant en calculant un score de clarté (Cronen-Townsend *et al.,* 2002). Ce score est en réalité la divergence de Kullback-Leibler entre le modèle de langue de l'ensemble de *hashtags* et le modèle de langue de la collection $\mathcal{C}$ d'articles Wikipédia :

$$\alpha = \sum_{w \in V} P(w|H_{\mathcal{T}}) \log \frac{P(w|H_{\mathcal{T}})}{P(w|\mathcal{C})}$$

où V représente le vocabulaire. Le modèle de langue des *hashtags* est estimé par retour de pertinence simulé :

$$P(w|H_{\mathcal{T}}) = \sum_{D \in R} P(w|D) P(D|H_{\mathcal{T}})$$

Nous utilisons pour cela une approche standard de retour de pertinence simulé. Celle-ci consiste à récupérer l'ensemble R constitué des 5 premiers documents de la collection $\mathcal{C}$ renvoyés pour la requête $H_{\mathcal{T}}$. Dans le modèle des *hashtags*, la probabilité $P(D|H_{\mathcal{T}})$ est estimée en appliquant le théorème de Bayes : $P(D|H_{\mathcal{T}}) = P(H_{\mathcal{T}}|D)P(D)$, où la probabilité $P(D)$ est égale à zéro pour les documents qui ne contiennent aucun mot de la requête. Plus les documents utilisés pour estimer le modèle de langue des *hashtags* sont homogènes, plus la divergence de Kullback-Leibler augmente. Ainsi le paramètre α permet de quantifier à quel point les *hashtags* sont précis et à quel point ils permettent de sélectionner des documents distincts du reste de la collection.

Seuls 23% des Tweets utilisés dans l'évaluation officielle de la tâche *Tweet Contextualization* d'INEX 2012 contiennent des *hashtags*. Lorsqu'il n'y en a pas, nous fixons logiquement $\alpha = 0$ dans l'équation 3.

3.3. *Génération des phrases candidates*

Pour un Tweet donné, nous sélectionnons les n articles Wikipédia les plus pertinents selon l'équation 3. Chaque article est découpé en phrases en utilisant la méthode PUNKT de détection de changement de phrases mise en œuvre dans `nltk`[4].

Dans ce travail nous fixons $n = 5$, et toutes les phrases des 5 premiers articles sont considérées comme des phrases candidates. Nous calculons ensuite différentes caractéristiques pour chacune de ces phrases qui nous permettrons de les classer et, ainsi, de former le contexte. Nous détaillons ces caractéristiques dans la section suivante.

4. Choix des phrases et formation du contexte

Pour pouvoir être compréhensible dans un cas d'utilisation mobile (sur un *smartphone* par exemple), le contexte doit avoir une taille limitée. Les recommandations de la tâche *Tweet Contextualization* d'INEX fixent la taille limite du contexte à 500 mots. Dans cette section, nous présentons la méthode que nous utilisons pour sélectionner les phrases candidates les plus pertinentes et générer le contexte.

4.1. *Caractéristiques des phrases*

Plusieurs caractéristiques entrent en compte lors de la sélection des phrases candidates. Ces dernières peuvent être regroupées en quatre catégories :

1) Importance de la phrase vis-à-vis du document d'où elle provient

2) Pertinence de la phrase par rapport au Tweet (y compris les *hashtags*)

4. `http://nltk.org/`

3) Pertinence de la phrase par rapport à une page web dont l'URL est dans le Tweet

4) Pertinence du document d'où provient la phrase par rapport au Tweet

Nous détaillons et justifions dans cette section le calcul des différentes caractéristiques que nous utilisons ensuite pour ordonner les phrases par importance et former le contexte. Nous rappelons quelques notations déjà utilisées dans cet article et nous en introduisons de nouvelles dans le tableau suivant :

$\mathcal{T}$	un tweet nettoyé
$H_{\mathcal{T}}$	les *hashtags* du Tweet $\mathcal{T}$
$U_{\mathcal{T}}$	l'URL présente dans le Tweet $\mathcal{T}$
S	une phrase candidate

Les caractéristiques décrites ci-dessous sont largement basées sur le calcul de mesures de recouvrement et de similarité cosine entre une phrase candidate $S = \{m_1, m_2, ..., m_i\}$ et un Tweet $\mathcal{T} = \{m_1, m_2, ..., m_j\}$. Soit $\mid \bullet \mid$ le cardinal de l'ensemble $\bullet$, le recouvrement en mots est donné par :

$$recouvrement(\mathcal{T}, S) = \frac{|S \cap \mathcal{T}|}{min(|S|, |\mathcal{T}|)}$$

Aussi, soient $\vec{S}$ et $\vec{\mathcal{T}}$ les représentations vectorielles de S et $\mathcal{T}$, et $||\vec{\bullet}||$ la norme du vecteur $\vec{\bullet}$, la similarité cosine est donnée par :

$$cosine(\mathcal{T}, S) = \frac{\vec{S} \cdot \vec{\mathcal{T}}}{\sqrt{||\vec{S}||\ ||\vec{\mathcal{T}}||}}$$

Les mesures décrites précédemment sont calculées à partir des représentations lexicales nettoyées des phrases et des Tweets. Nous supprimons les mots outils et appliquons le méthode de racinisation (*stemming*) des mots de Porter.

4.1.1. *Importance de la phrase dans le document*

L'importance d'une phrase par rapport au document dans lequel elle apparaît est estimée avec la méthode TextRank (Mihalcea, 2004). Chaque document est représenté sous la forme d'un graphe pondéré non dirigé G dans lequel les noeuds V correspondent aux phrases, et les arêtes E sont définies en fonction d'une mesure de similarité. Cette mesure détermine le nombre de mots communs entre les deux phrases, les mots outils ayant été au préalable supprimés et les mots restants *stemmés* avec l'algorithme de Porter. Pour éviter de favoriser les phrases longues, cette valeur est normalisée par les longueurs des phrases. Soit $\mathrm{freq}(m, S)$ la fréquence du mot m dans la phrase S, la similarité entre les phrases S_i et S_j est définie par :

$$\mathit{Sim}(S_i, S_j) = \frac{\sum_{m \in S_i, S_j} \mathrm{freq}(m, S_i) + \mathrm{freq}(m, S_j)}{\log(|S_i|) + \log(|S_j|)}$$

L'importance d'une phrase est évaluée en tenant compte de l'intégralité du graphe. Nous utilisons une adaptation de l'algorithme PAGERANK (Page *et al.,* 1999) qui inclut les poids des arêtes. Le score de chaque sommet V est calculé itérativement jusqu'à la convergence par :

$$p(V_i) = (1-d) + d \times \sum_{V_j \in \text{voisins}(V_i)} \frac{Sim(S_i, S_j)}{\sum_{V_k \in \text{voisins}(V_i)} Sim(S_k, S_i)} p(V_i)$$

où d est un « facteur d'amortissement » (typiquement dans l'intervalle $[0.8, 0.9]$) et $\text{voisins}(V_i)$ représente l'ensemble des nœuds connectés à V_i. Le score de la phrase S correspond au score du nœud qui la représente dans le graphe.

$$c_1 = p(S)$$

4.1.2. *Pertinence de la phrase par rapport au Tweet*

Intuitivement, les indicateurs de pertinence devraient être les plus importants pour sélectionner des phrases donnant des informations contextuelles par rapport au Tweet. Le recouvrement et la similarité cosine entre un Tweet $\mathcal{T}$ et une phrase candidate S sont les premières caractéristiques que nous avons mis en place.

$$c_2 = recouvrement(\mathcal{T}, S) \qquad c_3 = cosine(\mathcal{T}, S)$$

Tout en gardant la logique de l'utilisation des *hashtags*, nous calculons le recouvrement et la similarité cosine entre chaque phrase et l'ensemble des hashtags du Tweet.

$$c_4 = recouvrement(H_{\mathcal{T}}, S) \qquad c_5 = cosine(H_{\mathcal{T}}, S)$$

4.1.3. *Pertinence de la phrase par rapport à une page web*

Les Tweets contiennent parfois des URLs, liens pointant vers des pages web porteuses d'informations contextuelles. Nous utilisons le même type de mesure que précédemment et nous calculons ainsi le recouvrement et la similarité cosine entre une phrase candidate et le titre $titre(U_{\mathcal{T}})$ de la page web.

$$c_6 = recouvrement\big(titre(U_{\mathcal{T}}), S\big) \qquad c_7 = cosine\big(titre(U_{\mathcal{T}}), S\big)$$

De la même façon, nous calculons ces deux mesures entre le contenu entier $page(U_{\mathcal{T}})$ de la page web et une phrase candidate.

$$c_8 = recouvrement\big(page(U_{\mathcal{T}}), S\big) \qquad c_9 = cosine\big(page(U_{\mathcal{T}}), S\big)$$

4.1.4. *Pertinence du document par rapport au Tweet*

Les articles Wikipédia à partir desquels les phrases candidates sont extraites ont des importances contextuelles différentes par rapport à un Tweet donné. Ainsi, une phrase provenant d'un article bien classé a plus de chance d'être importante qu'une phrase provenant d'un article mal classé. Pour capturer ce comportement, nous définissons la dernière caractéristique comme étant le score d'un document par rapport à un Tweet et ses *hashtags*, normalisé sur l'ensemble R de tous les documents renvoyés :

$$c_{10} = \frac{s(\mathcal{T}, H_{\mathcal{T}}, D)}{\sum_{D' \in R} s(\mathcal{T}, H_{\mathcal{T}}, D')}$$

4.1.5. *Score final d'une phrase candidate*

Le score d'importance de chaque phrase candidate est obtenu par la combinaison linéaire des scores des critères présentés ci-dessus.

$$score = \sum_{x} \log(c_x + 1)$$

4.2. *Génération du contexte*

Le contexte d'un Tweet est généré par assemblage des phrases candidates les plus importantes. Il est cependant possible que le contexte ainsi obtenu contienne plusieurs phrases redondantes, ce qui dégrade à la fois sa lisibilité et son contenu informatif. Pour résoudre ce problème, nous ajoutons une étape supplémentaire lors de la génération des contextes.

Nous générons tous les contextes possibles à partir des combinaisons des N phrases ayant les meilleurs scores, en veillant à ce que le nombre total de mots soit optimal (i.e. en dessous du seuil de 500 mots et qu'il soit impossible d'ajouter une autre phrase sans dépasser ce seuil). La valeur N est fixée empiriquement au nombre minimum de phrases de meilleurs scores pour atteindre 500 mots, plus quatre phrases. Le contexte retenu au final est celui possédant le score global le plus élevé, ce score étant calculé comme le produit du score de la diversité du résumé, estimé par le nombre de n-grammes différents, et de la somme des scores des phrases.

Afin d'améliorer la lisibilité du contexte généré, si deux phrases sont extraites à partir d'un même document, l'ordre original du document est conservé.

5. Évaluation et discussion

Cette section débute par la description de la collection de test que nous utilisons. Nous présentons ensuite les résultats de notre méthode de contextualisation, et nous analysons l'importance des différentes caractéristiques dans le processus de sélection des phrases candidates.

5.1. *Cadre expérimental*

Nous utilisons la collection de test de la tâche *Tweet Contextualization* d'INEX 2012 pour nos expérimentations ainsi que les différentes données mises à disposition par les organisateurs (SanJuan *et al.,* 2012). La collection de documents Wikipédia est basée sur une capture de la version anglaise de l'encyclopédie en ligne datant de Novembre 2011 et comprend 3 691 092 articles. Nous avons indexé cette collection avec le moteur de recherche libre Indri [5] en supprimant les mots-outils présents dans la liste INQUERY. Une racinisation légère des mots est également appliquée par l'algorithme de Krovetz.

La collection de test comprend au total 1126 Tweets pour lesquels un système doit produire un contexte. Cependant, nous n'utilisons que le sous-ensemble de 63 Tweets pour lesquels des jugements de pertinence ont été réalisés. Ces jugements ont été générés par un processus de groupement des dix premières phrases des contextes de tous les participants qui ont ensuite été jugées manuellement par les organisateurs.

La mesure d'évaluation développée pour cette tâche ne prend pas en compte les exemples négatifs, seules les phrases jugées pertinentes ont été conservées. Les jugements sont donc un ensemble de phrases directement issues de Wikipédia et jugées pertinentes par les organisateurs en fonction de leur importance contextuelle par rapport à un Tweet. Certains Tweets peuvent ainsi avoir un contexte de référence composé d'un grand nombre de phrases, tandis que d'autres peuvent en avoir un nombre très réduit. Ces différences de taille ainsi que le fait qu'une seule référence soit disponible pour chaque Tweet empêchent l'utilisation de la mesure classique ROUGE (Lin, 2004) pour l'évaluation des contextes. Les organisateurs ont donc proposé une mesure d'évaluation qui calcule une divergence entre le contexte produit et les phrases jugées pertinentes (SanJuan *et al.,* 2012). Elle peut prendre en compte des unigrammes stricts, des bigrammes ou des bigrammes avec possibilité d'insertion. La mesure principale utilisée pour départager les systèmes est la troisième.

5.2. *Résultats de contextualisation*

Nous reportons dans le tableau 1 les résultats de contextualisation pour trois méthodes de recherche d'articles Wikipédia présentées dans la section 3 : l'approche standard par modèle de langue pour la RI (équation 1, notée **QL**), l'approche **MRF** (équation 2) et l'approche mixant MRF pour le Tweet et pour ses *hashtags* (équation 3, notée **MRFH**). Les scores étant calculés en tant que divergences, les scores les plus bas correspondent aux systèmes les plus performants.

Nous remarquons que les résultats sont relativement proches et qu'il n'y a pas de différence significative entre les trois approches. Néanmoins l'approche qui considère les *hashtags* dans la fonction de score des documents obtient les meilleurs résultats (avec $p = 0.17$ pour un t-test entre **QL** et **MRFH**). Les faibles différences observées

5. `http://www.lemurproject.org/indri.php`

	Unigrammes	**Bigrammes**	**Bigrammes à trous**
QL	0.7967	0.8923	0.8940
MRF	0.7883	0.8851	0.8865
MRFH	**0.7872**	**0.8815**	**0.8839**
1^{er} INEX 2012	0.7734	0.8616	0.8623

Tableau 1. *Résultats de contextualisation pour les 3 différents algorithmes de RI et l'ensemble des caractéristiques pour l'attribution des scores.*

entre les méthodes sont sans doute dues à la relative similarité entre les modèles de RI, même si l'on voit que l'utilisation de *hashtags* améliore sensiblement les scores. Il est néanmoins difficile de tirer des conclusions définitives étant donné que seuls 23% des Tweets utilisés pour l'évaluation contiennent au moins un *hashtag*. Nous reportons pour information les résultats officiels du meilleur système mais, à l'heure actuelle, leur approche n'est pas connue en détails. Grâce à l'analyse détaillée de l'influence des différentes caractéristiques proposée en section 5.3, nous avons pu établir que la borne supérieure de notre système était de 0.8824 ce qui est encore loin du meilleurs score. Cependant, il n'y a pas de différence statistiquement significative entre notre approche **MRFH** et le meilleur système d'INEX 2012.

Nous pensons que cette différence de score est due à deux biais lors de l'évaluation. Le premier se situe lors de la constitution des jugements : pour chaque Tweet, uniquement les dix premières phrases de chaque système sont considérées pour être ensuite jugées manuellement. Or, un des buts de cette tâche étant la lisibilité, les phrases les plus informatives ne se trouvent pas forcément en début de contexte pour pouvoir favoriser la cohérence globale et l'enchaînement des phrases. Le deuxième biais se situe au sein de la mesure d'évaluation elle-même. En effet, elle ne possède pas de composante visant à pénaliser les phrases non pertinentes. Ainsi, remplir le contexte avec des phrases très diverses permettra toujours d'obtenir des meilleurs scores que de faire attention et de ne pas ajouter de phrases dégradant la cohérence du contexte.

Pour illustrer ces biais, nous présentons dans la figure 3 un Tweet ainsi que le contexte produit par notre méthode ; qui a obtenu un score nul lors de notre évaluation. Or, même si ce contexte n'est à l'évidence pas parfait, il apporte tout de même des informations contextuelles sur le Tweet. On peut en effet apprendre que Van Gogh était un peintre et que "The Starry Night" est une de ses compositions, et dont le style transparaît sur d'autres de ses peintures.

Il est à noter que si les résultats obtenus par notre méthode lors de la campagne INEX ne sont pas les meilleurs, notre approche est celle qui apporte le meilleur compromis entre informativité et lisibilité. Néanmoins l'évaluation de lisibilité, qui a été faite manuellement, n'est pas reproductible et le travail présenté dans cet article est différent de celui réalisé pour INEX, nous ne pouvons donc pas reporter de résultats.

> « Very cool ! An interactive animation of van Gogh's "The Starry Night" http ://t.co/ErJCPObh (thanks @juliaxgulia) »

> Vincent van Gogh painted at least 18 paintings of "olive trees", mostly in Saint-Rémy in 1889. The olive tree paintings had special significance for Van Gogh. One painting, "Olive Trees in a Mountainous Landscape (with the Alpilles in the Background)", a complement to "The Starry Night", symbolized the divine. In both "The Starry Night" and his olive tree paintings, Van Gogh used the intense blue of the sky to symbolize the "divine and infinite presence" of Jesus. ...

Figure 3. *Les premières phrases d'un contexte produit par notre méthode. La mesure d'évaluation a attribué un score nul à ce contexte.*

5.3. *Importance des différentes caractéristiques*

En l'état, les caractéristiques calculées pour chacune des phrases candidates ont toutes la même importance dans le score final attribué à une phrase. Étant donné que les Tweets proposés pour la tâche QA@INEX 2011 n'avaient ni *hashtags* ni URL, nous n'avons pas pu entraîner notre système pour qu'il apprenne les poids de ces caractéristiques. Nous proposons néanmoins une analyse de leur importance sur les Tweets de l'année 2012. Bien évidemment, les chiffres présentés ici ne nous ont pas servi à paramétrer notre système, et les résultats présentés dans la section précédente ne tiennent pas compte de ces poids.

En principe, nous pourrions utiliser n'importe quelle méthode d'apprentissage pour apprendre les poids optimaux. Ici, nous utilisons un modèle de régression logistique. Ainsi, nous calculons toutes les caractéristiques présentées dans la section 4.1 pour chacune des phrases extraites et nous les lions à leur pertinence $r \in \{0, 1\}$. La variable r peut ainsi être vue comme une mesure de la contribution totale de toutes les caractéristiques utilisées dans le modèle et est habituellement définie comme $r = \bar{w}\bar{x}$. Spécifiquement, $\bar{x}$ est un vecteur de valeurs numériques représentant les caractéristiques, et $\bar{w}$ représente l'ensemble des poids relatifs de chaque caractéristique. Un poids positif signifie que la caractéristique correspondante améliore la probabilité d'obtenir r, un poids négatif signifie qu'elle la dégrade.

Nous pouvons observer dans le tableau 2 que les caractéristiques les plus significatives pour estimer la pertinence d'une phrase ne sont pas ou peu liées au Tweet. En effet, le TextRank ne concerne que l'importance de la phrase par rapport aux autres phrases du document, et le score du document est un score global. Le Tweet n'intervient dans ces cas qu'au moment de la recherche des articles. Comme on aurait pu s'y attendre, le recouvrement et la similarité cosine entre le Tweet et une phrase sont également des marqueurs de pertinence. Étonnamment, les *hashtags* ont une influence parfois négative et généralement aléatoire, tout comme les titres des pages web pointées par les URLs. Mais comme nous l'avons dit dans la section précédente,

Caractéristique	Nom	Valeur	Significativité
$c1$	**TextRank**	8.996	$p < 2^{-16}$
$c2$	**Recouvrement Tweet**	2.496	$p = 2.38^{-6}$
$c3$	**Cosine Tweet**	5.849	$p = 4^{-15}$
$c4$	Recouvrement *hashtags*	-2.051	$p = 0.1368$
$c5$	Cosine *hashtags*	0.671	$p = 0.3074$
$c6$	Recouvrement titre URL	1.373	$p = 0.2719$
$c7$	Cosine titre URL	0.788	$p = 0.6287$
$c8$	Recouvrement page URL	0.543	$p = 0.4337$
$c9$	**Cosine page URL**	10.374	$p = 0.0195$
$c10$	**Score document**	0.782	$p < 2^{-16}$

Tableau 2. *Valeurs optimales des poids des caractéristiques calculées pour les phrases candidates.*

les *hashtags* sont très peu nombreux dans les Tweets utilisés pour l'évaluation, ce qui peut expliquer ce comportement aléatoire. Enfin, seule la similarité cosine entre une phrase et le contenu d'une page web semble être faiblement significative.

Globalement, une phrase apporte des informations contextuelles par rapport à un Tweet si elle contient les mêmes mots que celui-ci, si elle apparaît dans un document pertinent, et si elle fait partie des phrases les plus importantes de ce dernier.

6. Conclusion

Nous avons présenté dans cet article une première approche pour la contextualisation de messages courts. Celle-ci se fait dans le cadre de la tâche Tweet Contextualization d'INEX et utilise Wikipédia comme corpus de référence pour la constitution des contextes. Les résultats de nos expériences suggèrent que l'utilisation des *hashtags* présents dans les Tweets aide à la recherche d'articles Wikipédia qui contiennent des phrases apportant des informations contextuelles. Nous avons également examiné l'influence de différentes caractéristiques calculées sur les phrases candidates ainsi que leur importance. Il apparaît que pour constituer un contexte, il est préférable de choisir les phrases les plus importantes des articles Wikipédia extraits. Les mesures de similarité entre les phrases et les Tweets sont également des indicateurs fiables, tandis que les *hashtags* semblent ici n'avoir qu'une influence aléatoire.

Une des limitations de notre approche est que le nombre d'articles Wikipédia utilisés pour extraire les phrases candidates est fixé manuellement. Idéalement, une méthode déterminant automatiquement ce nombre en fonction du Tweet permettrait de réduire le bruit et augmenterait indirectement la qualité des contextes générés. Même si l'utilisation du score du document permet de réduire l'effet de cette limitation, nous laissons cette amélioration pour des travaux futurs.

7. Bibliographie

Bakshy E., Rosenn I., Marlow C., Adamic L., « The role of social networks in information diffusion », *Proceedings of the 21st international conference on World Wide Web*, WWW '12, ACM, New York, NY, USA, p. 519-528, 2012.

Barzilay R., Elhadad M. et al., « Using lexical chains for text summarization », *Proceedings of the ACL workshop on intelligent scalable text summarization*, vol. 17, p. 10-17, 1997.

Boudin F., El-Bèze M., Torres-Moreno J.-M., « A Scalable MMR Approach to Sentence Scoring for Multi-Document Update Summarization », *Coling 2008 : Companion volume : Posters*, Coling 2008 Organizing Committee, Manchester, UK, p. 23-26, August, 2008.

Cronen-Townsend S., Croft W. B., « Quantifying query ambiguity », *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 104-109, 2002.

Dang H., « Overview of DUC 2005 », *Proceedings of the Document Understanding Conference*, 2005.

Lin C.-Y., « ROUGE : A Package for Automatic Evaluation of Summaries », *in* , S. S. Marie-Francine Moens (ed.), *Text Summarization Branches Out : Proceedings of the ACL-04 Workshop*, Association for Computational Linguistics, Barcelona, Spain, p. 74-81, July, 2004.

Metzler D., Croft W. B., « A Markov random field model for term dependencies », *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, ACM, New York, NY, USA, p. 472-479, 2005.

Mihalcea R., « Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization », *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Barcelona, Spain, p. 170-173, July, 2004.

Page L., Brin S., Motwani R., Winograd T., « The PageRank citation ranking : bringing order to the web. », 1999.

Radev D., Jing H., Styś M., Tam D., « Centroid-based summarization of multiple documents », *Information Processing & Management*, vol. 40, p. 919-938, 2004.

SanJuan E., Moriceau V., Tannier X., Bellot P., Mothe J., « Overview of the INEX 2012 Tweet Contextualization Track », *in* , P. Forner, , J. Karlgren, , C. Womser-Hacker (eds), *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

Segaran T., Hammerbacher J., *Beautiful Data : The Stories Behind Elegant Data Solutions*, O'Reilly Media, 2009.

Wei F., Li W., Lu Q., He Y., « Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization », *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, ACM, New York, NY, USA, p. 283-290, 2008.

Zhai C., Lafferty J., « A study of smoothing methods for language models applied to information retrieval », *ACM Transactions on Information Systems*, 2004.