Synthèse de flux de messages en temps réel

Abdelhamid Chellal, Mohand Boughanem, Bernard Dousset

Synthèse de flux de messages en temps réel

Abdelhamid Chellal — Mohand Boughanem — Bernard Dousset

IRIT, Université de Toulouse, CNRS, INPT, UPS, UT1, UT2J, France
118 route de Narbonne 31062 Toulouse Cedex 9, France
{abdelhamid.chellal, Mohand.Boughanem, Bernard.Dousset} @irit.fr

RÉSUMÉ. *La supervision d'événements à travers les réseaux sociaux ont connu un engouement certain ces dernières années. Cependant, le nombre considérable de messages publiés rend difficile, voire impossible, pour une personne de suivre "ce qui se passe autour de l'événement". Le filtrage et la génération en temps réel d'une synthèse de messages importants portant sur l'événement permet de répondre à ce besoin. La génération de cette synthèse pose plusieurs problèmes qui rend cette tâche difficile. En effet, la synthèse doit être concise, non redondante et couvrant le maximum de sous événements, et ce, en sus de la pertinence des informations qu'elle contient. Dans cet article, nous proposons une nouvelle approche de sélection des messages pour la génération, en temps réel, de résumés de flux de messages courts. La décision de sélectionner un message est prise instantanément, un nouveau message est ajouté au résumé si ses scores afférents à l'informativité et à la non redondance sont supérieurs à un seuil dynamique. L'approche proposée a été évaluée sur la collection TREC 2014 TS et elle a été comparée avec trois approches de l'état de l'art. Le résumé généré est de meilleure qualité que celui généré par les approches de "base" avec un nombre réduit de phrases.*

ABSTRACT. *Monitoring stream of social media posts has attracted significant attention in the last view years. Real time summarization in microblog aims at providing new relevant and non redundant information about an event as soon as it occurs. In this paper, we propose a novel continuous summarization approach in which information updates are pushed in real time. The decision to select/ignore incoming information is based on its informativeness and redundancy scores. The on-hand post is added to a summary only if the aforementioned scores are above a parametric free threshold. Our strategy was evaluated on a TREC Temporal Summarization 2014 data-set and it was compared with a well known baselines. The results reveal that our method outperforms all baselines and runs of the aforementioned task. The generated summaries are shorter and have higher precision than summaries generated by the baselines.*

MOTS-CLÉS : *filtrage temps réel, résumé temporal, informativité, redondance,*

KEYWORDS: *real time filtering, temproal summarization, informativeness, redundancy.*

1. Introduction

Partager des informations portant sur un événement est devenu une pratique très courante voire un réflexe dans les réseaux sociaux tels que Twitter. Les utilisateurs publient des informations actualisées sur l'évolution d'un événement comme un match de football ou une catastrophe naturelle. En effet, au 30 juin 2015, Twitter comptait 316 millions d'utilisateurs actifs par mois avec 500 millions de tweets envoyés par jour[1]. La supervision et le suivi d'événements à travers les réseaux sociaux peuvent être bénéfiques à plus d'un titre pour un journaliste ou une entreprise. A priori, tout le monde peut tirer profit de la disponibilité de cette masse d'information fraîche, diversifiée et presque "gratuite". Cependant, au regard de l'évolution exponentielle du nombre de messages publiés, il est très difficile pour un humain de suivre le flux d'information et construire un aperçu général sur l'évolution de l'événement objet d'intérêt, et ce, en particulier, lorsque il s'agit d'un événement qui dure dans le temps.

De plus, dans le cas d'événement inattendu à l'instar d'un attentat terroriste ou d'une catastrophe naturelle, l'utilisateur est à la recherche d'information actualisée et d'une mise à jour au fil de temps. Dans un tel scénario, une synthèse concise, non redondante mise à jour en temps réel qui met en évidence les sous événements les plus importants en présentant le développement de l'événement, serait très utile pour répondre au besoin d'information de l'utilisateur (Zubiaga *et al.*, 2012).

La tâche de filtrage et de synthèse en temps réel d'événement est différente de la recherche d'information classique dans laquelle les documents (ressources) sont triés en fonction de leur score de pertinence vis-à-vis de la requête. En effet, outre la pertinence qui est dans les deux tâches, le filtrage et la synthèse en temps réel exigent de retourner les nouvelles informations relatives à un événement dès qu'une information pertinente apparaît dans le flux des messages (Aslam *et al.*, 2014), et le résultat est une synthèse regroupant les messages pertinents non redondant.

Cette thématique de recherche a rencontré un réel engouement dans la communauté de Recherche d'Information ces dernières années (Sharifi *et al.*, 2010a), (Sharifi *et al.*, 2010b), (Ren *et al.*, 2013), (Xu *et al.*, 2013), (Olariu, 2014), (Mackie *et al.*, 2014). Récemment, une nouvelle tâche dédiée à cette thématique (Real-Time Filtering) a été introduite dans le cadre des compagnes TREC[2]. Le défi majeur sous-jacent à cette thématique porte sur le fait qu'en plus de la nécessite de retourner des tweets pertinents vis-à-vis de l'événement objet d'intérêt, la synthèse générée doit satisfaire les contraintes suivantes : éviter la redondance, être diversifiée en couvrant tous les sous-événements, et réduire la latence entre le temps de publication de l'information et le temps de notification. En effet, être en mesure de détecter l'information pertinente dés sa publication est très important dans le cadre de cette tâche, et ce, car l'information en tant que telle a une durée de vie, une date d'expiration, au-delà de laquelle elle n'est plus pertinente pour l'utilisateur. Ensuite, la redondance dans le

1. https ://fr.wikipedia.org/wiki/Twitter
2. https ://github.com/lintool/twitter-tools/wiki/TREC-2015-Track-Guidelines

flux de tweets est omniprésente vu que la même information est reliée par plusieurs utilisateurs à des moments différents.

Ainsi, une synthèse optimale devrait couvrir tous les sous événements sans redondance où chaque information est ajoutée dans la synthèse dès son apparition dans le flux. Il a été démontré dans (Shou *et al.*, 2013) que la synthèse dans les microblogs est un problème NP-difficile.

Par ailleurs, la synthèse (résumé) d'événements dans les flux des messages courts est différente de la synthèse (résumé) de documents, et ce pour les raisons suivantes : (i) les messages sont horodatés et l'information véhiculée est actualisée, elle représente une évolution temporelle de l'événement ; (ii) le flux de messages contient beaucoup de redondance, la même information est reliée par plusieurs utilisateurs sur des intervalles de temps différents (iii) la taille d'un message est généralement petite (limitée à 140 caractères dans le cas des tweets) ce qui fait que le contenu est souvent mal orthographié avec une utilisation fréquente des abréviations.

Dans cet article, nous présentons une nouvelle approche de génération de synthèse incrémentale de messages saillants (importants) portant sur un événement spécifié publié dans un flux de messages. Un message peut être soit une phrase extraite d'un document horodaté ou un blog soit un tweet. Il est considéré important s'il contient des informations nouvelles par rapport à celles déjà rencontrées dans le flux et non redondantes vis-à-vis des messages déjà sélectionnés dans la synthèse. L'importance d'un message est évaluée à travers trois critères, sa pertinence vis à vis l'événement recherché, son informativité représentée par la quantité d'informations apportées et estimée à travers l'entropie de Shannon (Shannon, 1948) et sa non redondance par rapport à la synthèse estimée, à travers la mesure de divergence Kullback-Leibler (Kullback et Leibler, 1951). Notre contribution porte sur :

1) La décision de sélectionner/ignorer un message est prise en temps réel. Cela permet d'éviter la mise en mémoire "buffering" du flux de messages et par la même de réduire le délai entre le temps de publication et le temps de notification à l'utilisateur ;

2) La sélection des messages est basée sur trois critères : la pertinence, l'informativité et la non redondance ;

3) Le seuil sur lequel se base la sélection des messages n'est pas un paramètre préfixé mais estimé d'une manière automatique.

Afin d'évaluer l'approche proposée, des expérimentations ont été réalisées sur la version filtrée de la collection TREC Temporal Summarization 2014 (TREC-TS-2014F)[3].

Cet article est organisé comme suit : Nous présentons dans la section 2 les travaux liés à la synthèse des microblogs et à la détection de la redondance. La section 3 décrit l'approche proposée. Dans la section 4, nous présentons et discutons les résultats de notre approche et nous les comparons aux résultats des trois approches de l'état de

3. http ://s3.amazonaws.com/aws-publicdatasets/trec/ts/index.html

l'art et aux systèmes ayant participé à la tâche TREC TS 2014. Nous terminons par une conclusion dans la section 5.

2. Etat de l'art

Dans cette section, nous présentons un bref aperçu sur les travaux relatifs à la génération automatique du résumé dans les microblogs suivi des approches de détection de la redondance en recherche d'information.

2.1. *Résumé automatique dans les microblogs*

D'une manière générale, les techniques du résumé automatique des documents peuvent être regroupées en deux catégories : abstractives et extractives (sélectives) (Ren *et al.*, 2013). Dans la première catégorie, le résumé généré contient des phrases, construites par le système de résumé automatique, qui n'apparaissent pas forcément dans les documents originaux. Tandis que dans la seconde catégorie, le résumé est construit à partir des phrases importantes sélectionnées directement des documents originaux. Pour cela, une fonction d'évaluation de l'importance des phrases est définie et les Top-K phrases sont sélectionnées. Ces deux stratégies ont été adoptées dans la littérature pour la génération automatique de résumé dans les microblogs. En ce qui concerne l'approche proposée dans cet article, elle appartient à la seconde catégorie. Ce choix d'adopter une approche sélective est motivé par le fait que c'est la stratégie la mieux adaptée à un scénario temps réel dans lequel le résumé est construit d'une manière incrémentale.

Les méthodes abstractives : La majorité des approches abstractives proposées dans la littérature se basent sur les graphes où les nœuds correspondent aux termes des tweets et les arrêtes modélisent leur ordre d'adjacence. La première approche appartenant à cet axe, est l'algorithme Phrase Reinforcement (PR) proposé par (Sharifi *et al.*, 2010b) qui génère une seule phrase pour résumer un événement. Dans cette approche un graphe ayant comme racine les mots clés de la requête est construit à partir des termes qui apparaissent dans les tweets. Un poids est attribué à chaque nœud (terme) en fonction de sa distance par rapport à la racine et sa fréquence d'occurrence dans le flux. Le chemin ayant le poids le plus élevé constitue la phrase "résumé".

Dans (Olariu, 2013) les auteurs proposent l'approche *Multi Sentence Compression* (MSC) qui représente les tweets collectés par un graphe orienté où chaque mot est représenté par un nœud et les arrêtes modélisent la relation d'adjacence entre les termes (bigramme). A chaque arrête est associée un poids calculé en fonction des occurrences des termes dans le flux des tweets. Le résumé est donné par le chemin ayant la plus petite moyenne des poids des arrêtes et dont la taille (nombre de mots) est supérieure à un certain seuil. Ainsi dans cette approche, le résumé généré contient des mots bigrammes peu fréquents dans le flux. Dans (Olariu, 2014) une extension du modèle précédent a été introduite dans laquelle le graphe est construit à partir des occur-

rences des trigrammes au lieu des bigrammes. Chaque nœud représente un bigramme et une arrête est ajoutée entre deux nœuds si les deux bigrammes afférents ont un terme en commun. A titre d'exemple, le graphe correspondant au tweet (w_1, w_2, w_3) est $(Debut, w_1) \rightarrow (w_1, w_2), (w_1, w_2) \rightarrow (w_2, w_3)$ et $(w_2, w_3) \rightarrow (w_3, Fin)$. Les poids des nœuds et des arrêtes correspondent aux occurrences des bigrammes et des trigrammes respectivement. Dans cette approche, les top-k nœuds ayant le poids le plus fort sont sélectionnés comme noyau. Ensuite pour chaque nœud sélectionné, une phrase "résumé" est construite à partir du chemin qui contient ce nœud et qui maximise un score d'importance estimé à partir les poids des nœuds et des arrêtes ainsi que la fréquences des termes. L'inconvénient majeur de cette approche est que l'utilisation de trigramme augmente de manière significative le nombre de nœuds dans le graphe.

Les méthodes sélectives : Concernant la synthèse extractive, les approches proposées peuvent être classées en deux catégories : Dans la première catégorie, les tweets sont modélisés dans des graphes où les nœuds représentent les tweets et les arrêtes correspondent à la similarité entre deux tweets. Pour chaque arrête un poids relatif au degré de similarité entre les tweets afférents est calculé. Dans (Liu *et al.*, 2012), d'autres caractéristiques liées au réseau social comme le nombre de retweets et la lisibilité du texte sont combinés avec la mesure de similarité pour évaluer le poids des arrêtes . Un score d'importance est calculé pour chaque nœud en fonction de sa similarité par rapport aux autres nœuds. Ainsi, le résumé est construit à partir des nœuds ayant le plus grand score d'importance.

Dans la seconde catégorie basée sur les caractéristiques des tweets, une variété de caractéristiques textuelles a été exploitée telles que le modèle de langue (O'Connor *et al.*, 2010), la fréquence des termes (Liu *et al.*, 2012) et TF-IDF (Chakrabarti et Punera, 2011). Dans certaines approches, la problématique de synthèse sélective de tweets a été formalisée sous forme d'un problème d'optimisation combiné avec des techniques de classification. Dans (Liu *et al.*, 2011) la synthèse automatique de tweets a été modélisée sous la forme du programme linéaire suivant : Pour un sujet donné, if faut sélectionner un nombre k de tweets qui maximise la somme des poids des concepts (les n-gramme les plus fréquents) sous deux contraintes : la première contrainte porte sur le nombre de tweets à sélectionner et la seconde concerne le nombre de termes total dans le résumé.

Dans Zubiaga et al. (2012) l'importance des tweets est évaluée en fonction des fréquences de ses termes et sa non redondance par rapport aux tweets déjà sélectionnés dans la synthèse. La non redondance est estimée par une divergence de Kullback-Leibler (Kullback et Leibler, 1951). Il s'agit de la première approche qui propose la génération en temps réel de résumé dans les microblogs. Cependant, elle est dédiée aux événements programmés comme les matchs de football.

Dans (Sharifi *et al.*, 2010c) les auteurs ont introduit l'approche hybrideTF-IDF dans laquelle les tweets sont triés en fonction de leur score de pertinence estimé à travers la moyenne des poids de leurs termes. Le poids d'un terme est évalué par la fonction hybrideTF-IDF. Les Top-k tweets sont sélectionnés en excluant ceux qui ont une similarité par rapport aux tweets déjà sélectionnés supérieure à un seuil prédéfini.

Dans ce contexte, la fonction cosinus a été utilisée pour estimer la similarité entre les tweets. Dans cette approche, les auteurs suggèrent de considérer la collection de tweets comme un seul document pour calculer la fréquence du terme TF. L'approche Sumbasic (Nenkova et Vanderwende, 2005) initialement proposée pour le résumé des documents s'est avérée aussi efficace pour les microblogs selon (Karkali *et al.*, 2014). L'intuition sous-jacente à cette approche est que les mots les plus fréquents dans les documents ont une probabilité plus élevée d'être sélectionnés dans le résumé que les mots les moins fréquents.

La première comparaison en termes de performance absolue des différentes approches de synthèse automatique des microblogs, a été réalisée par Inouye et Kalita, (2011). Cette étude a fait ressortir que les méthodes basées sur la fréquence des termes obtiennent des performances meilleures, et ce, au regard de la spécificité de tweets (non structurés et courts). Dans ce contexte, il a été rapporté que l'approche hybrideTF-IDF est l'une des meilleures approches. Récemment, dans (Mackie *et al.*, 2014) les auteurs ont comparé onze approches de génération automatique de synthèse en utilisant quatre corpus de microblog. Les résultats révèlent que SumBasic (Nenkova et Vanderwende, 2005) et hybrideTF-IDF (Sharifi *et al.*, 2010c) sont les plus efficientes. Partant de ce résultat et conforment aux recommandations des auteurs, nous avons adopté lesdites méthodes comme ligne de base dans l'évaluation expérimentale de notre approche.

Notre travail s'inscrit dans la ligne de recherche portant sur l'adoption de la stratégie sélective pour la construction de résumé en exploitant des statistiques textuelles. Notre approche diffère des approches citées ci-dessus par : (i) Le résumé est construit d'une manière incrémentale où la décision de sélectionner / ignorer un message est prise instantanément en temps réel et sans utilisation de connaissances externes. (ii) En outre de la pertinence, la sélection est basée sur deux critères (informativité et non redondance) combinées sous forme de contraintes conjonctives. (iii) Dans notre approche, le seuil de sélection est estimé d'une manière automatique. Noter aussi, que l'approche proposée est indépendante du type d'événement alors que l'approche proposée dans (Olariu, 2014) est applicable uniquement pour des événements planifiés.

2.2. *Détection de la redondance*

La détection de redondance repose sur des mesures de similarité / divergence comme la distance de Manhattan, la fonction cosinus ou la comparaison des distributions des probabilités. Elle est habituellement utilisée avec la nouveauté dans les travaux relatifs à la détection et le suivi des topics (Topic Detection et Tracking) (Markou et Singh, 2003). Deux classes d'approche peuvent être distinguées. Dans la première, dite document-document, le nouveau document est comparé à tous ceux de la collection. L'inconvénient majeur de ce type d'approche réside dans la complexité des calculs dans le cas de collections volumineuses. Dans la seconde classe, et pour diminuer cette complexité, un nombre réduit de documents résumant la collection est généré, la similarité du nouveau document est alors mesurée uniquement avec le sous

ensemble de documents (Karkali *et al.*, 2014). Notre approche se base sur cette seconde classe. Ce choix est motivé par la nécessité de réduire la complexité du calcul dans le traitement temps réel du flux.

3. Génération en temps réel de résumé basée sur l'informativité et la redondance

L'approche que nous proposons traite en temps réel le flux des messages puis décide de garder ou d'ignorer le message en cours en fonction de plusieurs critères. Les messages sont traités selon leur ordre chronologique. Le problème de génération en temps réel de synthèse peut être défini comme suit :

1) $\forall T_i, T_j \in R,\ t_i < t_j$. où R est le résumé et t_i , t_j sont les temps de publication des messages T_i et T_j respectivement. Autrement dit, les messages sont collectés et traités selon leur ordre chronologique ;

2) $\forall T_i \in R,\ \Delta t = \tau_i - t_i \simeq 0$, où τ_i est le temps de notification (temps de prise de décision de sélectionner le message) ;

3) $\forall T_i, T_j \in R,\ T_i \nsim T_j$; c'est à dire que les deux messages T_i et T_j apportent des informations significativement différentes afin d'éviter la redondance dans la synthèse et d'assurer une couverture maximale de tous les sous-événements ;

4) $R \prec R'$, le résumé R est préféré par rapport au résumé R' si R couvre au moins le même nombre de sous-événements (aspects) que R' avec un nombre réduit de messages (résumé concis).

Ces critères sont satisfaits dans l'approche proposée comme suit : (i) La décision de sélectionner/ignorer un message est prise en temps réel dès la réception d'un message. (ii) Le processus de prise de décision se base sur trois dimensions : la pertinence, l'informativité et la redondance. La première dimension permet de réduire le bruit et de limiter le traitement aux messages potentiellement importants. La seconde dimension détecte les messages qui apportent une quantité d'information significative par rapport à ce qu'il a été déjà vu dans le flux. Tandis que la troisièmement dimension permet d'éviter de sélectionner des messages similaires à ceux déjà ajoutés dans le résumé ce qui permet de réduire la redondance.

Autrement dit, un nouveau message candidat est ajouté au résumé si et seulement si ses scores afférents à l'informativité et la non redondance sont supérieurs à un certain seuil. Cela nous ramène à la question suivante : comment définir les seuils de sélection ? En effet, les statistiques (fréquences des termes) varient dans le temps ce qui rend difficile la définition à priori des seuils de sélection. En outre, apprendre ces seuils à partir d'une collection d'entrainement pose le problème de dépendance vis à vis aux types d'événements de la collection d'entrainement. Pour remédier à cette problématique, nous proposons de fixer ces seuils d'une manière adaptative en fonction des valeurs précédemment observées.

3.1. *Filtrage de messages*

Pour atteindre les exigences citées ci-dessus, notre approche agit comme un filtre à trois niveaux. Le premier filtre porte sur la pertinence vis à vis la requête. Tous les messages qui ne contiennent pas au moins deux mots de la requête sont ignorés. les tweets qui passent ce premier filtre sont considérés comme des tweets candidats à la sélection dans le résumé. Les deux filtres suivants portent respectivement sur l'informativité et la non redondance. Si on considère que R^{t_i} est l'ensemble des messages résumant un événement à l'instant t_i, un nouveau message candidat T_i sera ajouté à R^{t_i} si et seulement si :

$$\begin{cases} IS(T_i) >= \max\limits_{\forall T_j \in F^{t_i},\ t_i < t_j} [IS(T_j)] \\ RS(T_i) >= \max\limits_{\forall T_j \in R^{t_i},\ t_i < t_j} [RS(T_j)] \end{cases} \quad [1]$$

Où $IS(T_i)$ et $RS(T_i)$ sont les scores d'informativité et de non-redondance de message T_i respectivement. F^{t_i} et R^{t_i} sont le flux de messages et le résumé à l'instant t_i (temps de publication du message T_i).

Ainsi T_i est ajouté au résumé R^{t_i} si ses scores d'informativité et de non-redondance sont respectivement supérieurs à ceux obtenus par tous les messages du flux (pour l'informativité) et ceux du résumé (pour la non redondance).

D'autres options auraient pu être considérées pour combiner les deux critères. En particulier, on aurait pu combiner linéairement les deux scores ; on aurait eu un seul seuil à contrôler. Cette alternative n'a pas été considérée car elle pourrait conduire à l'ajout dans le résumé des messages redondants ayant une informativité élevé où bien des messages non redondants mais peu informatifs. Nous proposons la combinaison de ces deux critères sous forme d'une condition conjonctive qui permet d'avoir une complémentarité entre eux.

3.2. *Score d'informativité d'un message*

En théorie de l'information, la quantité d'information apportée par un message est évaluée à travers l'entropie de Shannon (Shannon, 1948). Pour évaluer l'importance d'un message, nous proposons l'utilisation de cette mesure afin d'estimer son informativité par rapport au flux d'information. L'intuition sous-jacente à cette proposition est qu'un nouveau message est considéré comme pertinent (peut d'être ajouté au résumé) s'il apporte une information différente où complémentaire par rapport aux précédents messages dans le flux. L'informativité, plus précisément la quantité d'information, apportée par le message, mesurée par son entropie, est supérieure à la quantité d'infor-

mation apportée par ses prédécesseurs. Ainsi, le score d'informativité d'un message T publié à l'instant t est mesuré comme suit :

$$IS(T) = -\sum_{w_i \in T} P^t(w_i) \times log_2(P^t(w_i)) \quad [2]$$

Où $P^t(w_i)$ représente la probabilité d'occurrence de terme w_i à l'instant t (timestamps de la phrase T) dans le flux. Cette probabilité est estimée comme suit :

$$P^t(w_i) = \frac{\#Messages\ ou\ w_i\ apparait\ à\ l'instant\ t}{\#Messages\ dans F^t} \quad [3]$$

La probabilité $P^t(w_i)$ peut être vue comme une DF (Document Frequency) inverse où une fréquence documentaire relative. Il convient de noter que dans les microblogs, la fréquence d'occurrence d'un mot est généralement égale à 1 où 0. Ainsi, pour un terme donné, sa fréquence d'occurrence (TF) dans le flux (la collection) est égale au nombre de messages où il apparait (fréquence documentaire DF). De ce constat, il ressort que la notion de TF est peu pertinente.

Un message long contenant des termes importants sera préféré par rapport un message court. A titre d'exemple, supposons que $R = \{t_1 = \{w_1, w_2\}\}$ et que nous avons trois nouveaux messages en entrée $t_2 = \{w_1, w_3\}$, $t_3 = \{w_4, w_5\}$ et $t_4 = \{w_2, w_3, w_4, w_5\}$. Le message t_4 obtiendra les scores d'informativité et de non-redondance les plus élevés.

3.3. *Score de non redondance*

Pour évaluer la divergence entre deux messages, nous proposons de mesurer l'écart entre leur modèle de langue respective en utilisant la divergence de Kullback-Leibler (KL) (Kullback et Leibler, 1951) définie comme suit :

$$KL(T, T') = \sum_{w_i \in T \cap T'} \theta_T(w_i) \log \frac{\theta_T(w_i)}{\theta_{T'}(w_i)} \quad [4]$$

où θ_T est le modèle de langue uni-gramme de message T et $\theta_T(w_i)$ est la probabilité d'occurrence de terme w_i dans le message T.

Pour évaluer le score de non redondance d'un message candidat ayant passé les deux premiers filtres relatifs à la pertinence et l'informativité, nous proposons de mesurer la non redondance vis à vis aux messages déjà sélectionnés dans le résumé. Les scores de divergence entre le nouveau message et tous les messages du résumé actuel peuvent être agrégés dans un seul score global. Cependant, un message peut avoir une grande divergence moyenne tout en être très proche à un message particulier dans

le résumé. Pour cela, nous proposons de considérer le score de divergence minimum comme score de non redondance du nouveau message.

Plus précisément, ce que nous proposons est de considérer la divergence entre le nouveau message et le message le plus proche parmi les messages déjà sélectionnés. Le message le plus proche (le moins divergeant) est celui dont la divergence est la plus petite. Ainsi, le score de non redondance d'un message vis à vis du résumé est défini par le minimum de divergence KL obtenue par rapport à tous les messages du résumé actuel (à l'instant t_i)R^{t_i} tel que mentionné par la formule suivante :

$$RS(T_i) = \min_{\forall T_j \in R^{t_i}} KL(T_i, T_j) \quad [5]$$

Pour remédier au problème de probabilité nulle, nous pouvons lisser les probabilités en utilisant le lissage de Jelinek-Mercer (JM) ou de Dirichlet (D). Le résultat de nos expérimentations révèle que le lissage de Dirichlet convient mieux au contexte de synthèse temps réel des documents courts. Ce dernier est défini par la formule suivante :

$$\theta_T(w_i) = \frac{TF_t(w_i) + \mu P_F^t(w_i)}{|T| + \mu} \quad [6]$$

Où $TF_t(w_i)$ est la fréquence d'occurrence du mot w_i dans le message T et μ est le paramètre de lissage. $P_F^t(w_i)$ est la probabilité d'occurrence de terme w_i dans le flux F à l'instant t. Elle est évaluée en utilisant l'estimation du maximum de vraisemblance (ML), et ce, comme suit :

$$P_F^{t_i}(w_i) = \frac{\#(w_i)\ dans\ F^{t_i}}{|F^{t_i}|} \quad [7]$$

Dans nos expérimentation, le paramètre de lissage μ a été fixé à 1000. Cette valeur a été estimée d'une manière expérimentale en variant μ de 10 à 2000 par un pas de 100.

4. Évaluation expérimentale

Afin de valider notre approche, une série d'expérimentations a été effectuée sur la collectionne TREC Temporal summarization 2014. Trois objectifs principaux sont identifiés :

1) Évaluer l'impact des critères relatifs à l'informativité et non redondance : combinés où utiliser séparément ;

2) Évaluer l'impact du seuil dans l'équation [1] ;

3) Comparer les performances de l'approche proposée avec ceux de l'état de l'art et ceux obtenus dans TREC TS 2014.

A cet effet, deux versions de l'approche proposée ont été testées. Dans la première, les deux filtres relatifs à l'informativité et la redondance sont appliqués dans le processus de fabrication du résumé. Tandis que dans la seconde version, un seul filtre est appliqué à la fois. De plus, deux manières différentes d'estimer le seuil de sélection (équation 1) ont été testées, la moyenne et le maximum des scores précédemment observés.

Les résultats obtenus par notre approche ont été confrontés à ceux de trois approches de l'état de l'art. La première est l'approche SumBasic (Nenkova et Vanderwende, 2005) recommandée pour être adoptée comme ligne de base par (Mackie *et al.*, 2014). La deuxième ligne de base est l'approche hybrideTF-IDF (Sharifi *et al.*, 2010c) qui a été rapporté comme étant une des meilleures approches de résumé dans les microblogs. Aussi, nous avons implémenté l'approche TF-IDF combinée avec la fonction cosinus pour l'estimation du score de non redondance. Il convient de signaler que ces lignes de bases ont été ajustées pour correspondre au traitement en temps réel et en utilisant le même seuil de sélection adopté par notre approche (équation N°1). Les équations ci-dessous décrivent les fonctions hybrideTF-IDF et Sumbasic pour un message donné T, respectivement :

$$HybridTF - IDF(T) = \frac{\sum_{w_i \in T} TF(w_i) \times IDF(w_i)}{\max[Minimum\ seuil, |T|]} \quad [8]$$

$$TF(w_i) = \frac{\#(w_i)\ dans\ F}{\#mot\ dansF},\ IDF(w_i) = log_2(\frac{\#Messages}{\#Messages\ ou\ w_i\ apparait}) \quad [9]$$

$$Sumbasic(T) = \sum_{w_i \in T} \frac{P(w_i)}{|T|},\ P(w_i) = \frac{\#w_i\ dans\ F}{\#T\ ou\ w_i\ apparait} \quad [10]$$

4.1. *La collection des données*

Les expérimentations ont été menées sur la collection des documents TREC *Temporelle Summarization* (TS) 2014 qui contient des documents horodatées en anglais issus de différentes sources du Web (presse, blog, réseaux sociaux, etc.) et collectés durant la période allant d'octobre 2011 à avril 2013. Le but de cette tâche est d'évaluer les approches de supervision des événements en se focalisant sur la détection le plutôt possible des nouvelles informations publiées dans un flux. Dans nos expérimentations, nous avons utilisé la version filtrée de la collection (TREC-TS-2014F) qui regroupe les documents susceptibles de contenir au moins une phrase pertinente. Pour l'année 2014, quinze (15) sujets ont été définis par les organisateurs de TREC. En plus, les fragments d'informations vitaux (pertinent et apporte une information nouvelle), qu'un résumé doit contenir, ont été fournis pour chaque sujet. Une phrase d'un document est jugée pertinente si elle est associée à au moins une information vitale. Les documents sont traités par ordre chronologique et le système de génération de résumé enregistre le temps de notification par rapport au flux de document au moment de la sélection d'une phrase dans le résumé. Les messages traités sont soit des tweets soit

Tableau 1. *Comparaison entre les différentes configurations de l'approche proposée.*

Metrique	H	nEG(S)	C(S)	Nb de phrases
KL-Entropie-MAX	**0.1449**	**0.0793**	0.3167	171
Entropie-Only-MAX	0.1191	0.0562	0.3925	333
KL-Entropie-AVG	0.0878	0.0383	0.4151	758
KL-Only-MAX	0.0846	0.0382	0.3427	477
Entropie-Only-AVG	0.0606	0.0239	0.528	1812
KL-Only-AVG	0.0179	0.0064	**0.5400**	7986

des phrases extraites des documents horodatés. Dans l'approche décrite dans cet article, la décision de sélectionner/ignorer une phrase est prise immédiatement. Ainsi, le temps de notification correspond au temps de publication du document.

4.2. *Métriques d'évaluation*

Dans la tâche TREC TS, les performances sont évaluées en fonction de la pertinence et la couverture lesquelles sont estimées à travers les métriques suivantes : (i) le gain escompté normalisé (nEG) qui permet d'évaluer la pertinence des phrases sélectionnées dans le résumé. (ii) la compréhension "Comprehensiveness" (C) qui estime la couverture du résumé. Ces métriques sont similaires aux mesures de précision et rappel respectivement (Aslam *et al.*, 2014), mais en pénalisant la redondance et la latence lors de la détection d'informations. (iii) la moyenne harmonique (H) entre nEG et C a été définie comme étant la métrique officielle. Les résultats sont classés en fonction de cette mesure. Tous les détails concernant ces métriques sont disponibles sur http ://www.trec-ts.org/downloads.

4.3. *Résultats*

Dans nos expérimentations, nous avons évalué plusieurs configurations de l'approche proposée. Dans la première configuration notée (KL-Entropie), les deux filtres (informativité et non-redondance) ont été utilisés dans le processus de sélection des phrases tandis que dans la seconde configuration un seul critère a été utilisé (KL) ou (Entropie). Les suffixes "MAX" et "AVG" représentent le seuil utilisé (le maximum ou la moyenne des scores d'informativité et de non redondance respectivement).

Impact de la combinaison de deux dimensions : Le tableau 1 liste les performances des différentes configurations. En ressort que la combinaison des deux composants augmente de manière significative le gain (nEG) tandis que la couverture (C) est légèrement réduite (par rapport à la couverture obtenue lorsque les critères de sélectionne sont utilisés séparément) ce qui permet d'avoir une moyenne harmonique

(H) relativement élevée. Cette réduction du couverture dans la version KL-Entropie est attendue vu que l'application des deux filtres réduit le nombre de phrases dans le résumé généré. Toutes les phrases que contient le résumé généré en utilisant les deux filtres informativité (entropie) et non redondance (KL) sont incluses dans les résumés générés lorsque un seul filtre est utilisé $R(KL-Entropie) \subset R(Entropie-Only)$ et $R(KL-Entropie) \subset R(KL-Only)$. Ce résultat confirme notre intuition relative à la nécessité de combiner plusieurs dimensions pour la sélection de fragment d'information lors de la construction d'une synthèse en temps réel. De plus, il semble y avoir un compromis entre le gain (nEG) et la couverture (C) où il est plus difficile d'améliorer la couverture que le gain, et ce, au regard du fait que le résumé est construit d'une manière incrémentielle et il doit avoir une taille réduite.

En outre en comparant l'impact de chaque dimension seule, nous constatons que l'informativité joue un rôle plus important par rapport à la redondance. En effet, le filtrage selon le score d'informativité (entropie) permet d'avoir un résumé meilleur en terme de gain (nEG) et de couverture (ligne 2). Le score d'informativité représente la quantité d'information apportée par une phrase par rapport au flux. Ce score, tel que défini dans l'équation 2, augmente avec la fréquence des termes dans le flux. Ainsi une phrase est ajouté au résumé si elle contient des termes importants (fréquents) ce qui permet de sélectionner des phrases pertinentes. Ce résultat signifie que le filtre relatif à l'informativité pré-sélectionne les informations candidates et le filtre de non-redondance approuve ou désapprouve ce choix.

Impact de seuil : Le seuil de sélection dans les deux filtres (équation 1) porte sur la valeur maximale des valeurs précédentes. Nous avons testé la moyenne comme seuil pour étudier l'impact de ce paramètre. Le résultat obtenu (Tableau 1) montre que l'utilisation de la moyenne comme seuil donne de faible performance en termes de gain et de moyenne harmonique (H) entre (nEG) et la couverture. Cependant, il donne une couverture meilleure. Cela peut être expliqué par le nombre relativement élevé de phrases sélectionnées dans le résumé en raison du fait que l'utilisation de la moyenne comme seuil est moins restrictive par rapport au maximum. Le nombre moyen de phrases dans le résumé généré par le KL-Entropie-AVG est de 758 alors que en moyenne le nombre de phrases dans le résumé généré par KL-Entropie-MAX est de 171. Nous observons que la couverture est proportionnelle à la taille du résumé (nombre de phrase).

Comparaison des résultats par rapport à l'état de l'art : Dans le tableau 2, nous comparons les performances de notre approche par rapport aux résultats obtenus par les 3 approches de l'état l'art et aux résultats officiels de TREC TS 2014 en terme de gain (precision) et de couverture (rappel). Cette comparaison révèle que l'approche KL-Entropie donne une meilleure moyenne harmonique (H) entre le la couverture (C) et le gain (nEG). Elle surpasse l'approche qui a été classée la première dans TREC TS 2014 (2APSal) (Aslam *et al.*, 2014) de 24% et l'approche hybrideTF-IDF de 19%. Le gain de l'approche proposée est constamment supérieur par rapport aux autres approches tandis que la couverture est plus petite. Ce résultat peut être expliqué par le fait que la combinaison des deux scores dans une contrainte conjonctive et l'utilisation

Tableau 2. *Comparaison des performances de l'approche proposée par rapport aux lignes de bases et les résultats des participants à TREC TS 2014*

Metrique	H	nEG(S)	C(S)
KL-Entropy-MAX	**0.1449**	**0.0793**	0.3167
HybridTF-IDF	0.1207	0.0593	**0.3796**
SumBasic	0.0904	0.0402	0.3351
TF-IDF	0.0987	0.0483	0.3672
TREC TS 2014 task results			
2APSal	**0.1162**	0.0631	0.3220
Q1	0.1110	0.0657	**0.4088**
Q2	0.1091	0.0632	0.3979
Average all runs	0.0620	0.0327	0.3615

du maximum des valeurs précédentes comme seuil de sélection réduit le nombre de phrases sélectionnées ce qui augmente considérablement le gain. L'entropie donne un score élevé pour les phrases contenant des termes fréquents dans le flux et le filtre de redondance rejettera les phrases qui contiennent des mots fréquents dans le résumé. Ainsi, seules les phrases contenant un bon mélange de termes fréquents dans le flux et nouveaux par rapport au résumé seront sélectionnées.

Enfin, dans la figure 1, nous comparons la taille moyenne du résumé généré par chaque approche par rapport au nombre moyen d'informations vitales à retrouver pour les 15 topics. Nous observons que le résumé généré par l'approche proposée est plus petit (plus concis) que le résumé généré par les trois approches de l'état de l'art et l'approche 2APSal selon le résultat obtenus lors de la participation à TREC TS 2014 (Aslam *et al.*, 2014). La taille moyenne du résumé produit par KL-Entropie est de 171 alors qu'elle est de 381 et 340 pour 2APSal et hybrideTF-IDF respectivement. On note aussi que le nombre moyen d'informations vitales "golden nuggets" par topic est de 94. Ce résultat montre que notre approche génère un résumé concis ayant un gain meilleur avec une réduction de la couverture relativement peu significative.

5. Conclusion

La synthèse en temps réel d'événement est un sujet qui a connu récemment un engouement dans la recherche d'information. Dans cet article, nous avons proposé une nouvelle méthode de génération de résumé en temps réel. Les expérimentations ont été réalisées sur la collection de documents filtrés de la compagne TREC TS 2014. Les résultats révèlent que l'approche décrite donne des meilleures performances en terme de précision et de moyenne harmonique entre le rappel et la précision par rapport aux systèmes ayant participé à TREC TS 2014 et trois approches de l'état de l'art.

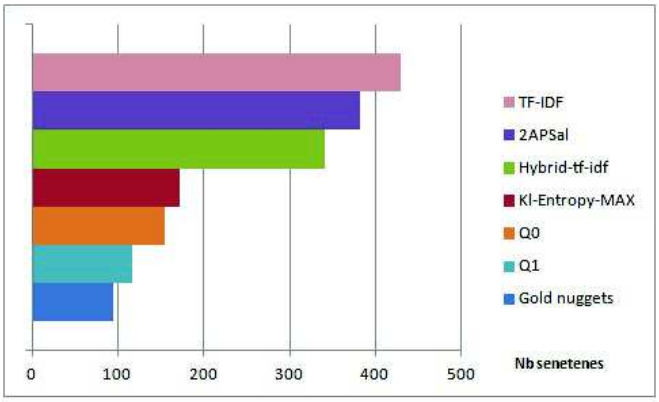

Figure 1. *Nombre moyen de phrases dans le résumé par sujet*

Notre approche est indépendante de l'événement, la génération de résumé est réalisée d'une manière incrémentale en temps réel au lieu de catégoriser les messages en sous-événements.

Ces résultats montrent que les statistiques textuelles sur le flux peuvent être exploités dans le filtrage et la génération de synthèse en temps réel. Le résumé obtenu est court et compact avec une bonne précision. Cependant la couverture diminue avec l'augmentation de la précision. Pour y remédier, d'autres pistes sont à explorer notamment en ce qui concerne le choix du seuil de sélection et l'identification d'autres dimensions tel que la diversité comme filtre à ajouter pour améliorer la qualité du résumé généré.

6. Bibliographie

Aslam J. A., Ekstrand-Abueg M., Pavlu V., Diaz F., McCreadie R., Sakai T., « TREC 2014 Temporal Summarization Track Overview », *Proceedings of The Twenty-Third Text REtrieval Conference, TREC, 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*, 2014.

Chakrabarti D., Punera K., « Event Summarization Using Tweets », *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.

Karkali M., Rousseau F., Ntoulas A., Vazirgiannis M., « Using temporal IDF for efficient novelty detection in text streams », *CoRR*, 2014.

Kullback S., Leibler R. A., « On Information and Sufficiency », *The Annals of Mathematical Statistics*, vol. 22, n° 1, p. 79-86, 03, 1951.

Liu F., Liu Y., Weng F., « Why is "SXSW" Trending ? : Exploring Multiple Text Sources for Twitter Topic Summarization », *Proceedings of the Workshop on Languages in Social Media*, LSM '11, p. 66-75, 2011.

Liu X., Li Y., Wei F., Zhou M., « Graph-Based Multi-Tweet Summarization using Social Signals », *Proceedings of COLING 2012*, The COLING 2012 Organizing Committee, Mumbai, India, p. 1699-1714, December, 2012.

Mackie S., McCreadie R., Macdonald C., Ounis I., « Comparing Algorithms for Microblog Summarisation », *Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings*, p. 153-159, 2014.

Markou M., Singh S., « Novelty Detection : A Review - Part 1 : Statistical Approaches », *Signal Process.*, vol. 83, nº 12, p. 2481-2497, December, 2003.

Nenkova A., Vanderwende L., The Impact of Frequency on Summarization, Technical Report nº MSR-TR-2005-101, MSR-TR-2005-101, January, 2005.

O'Connor B., Krieger M., Ahn D., « TweetMotif : Exploratory Search and Topic Summarization for Twitter », *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, 2010.

Olariu A., « Hierarchical Clustering in Improving Microblog Stream Summarization », *Computational Linguistics and Intelligent Text Processing - 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II*, p. 424-435, 2013.

Olariu A., « Efficient Online Summarization of Microblogging Streams », *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2 : Short Papers*, Association for Computational Linguistics, Gothenburg, Sweden, p. 236-240, April, 2014.

Ren Z., Liang S., Meij E., de Rijke M., « Personalized Time-aware Tweets Summarization », *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, ACM, New York, NY, USA, p. 513-522, 2013.

Shannon C. E., « A Mathematical Theory of Communication », *Bell System Technical Journal*, vol. 27, p. 379-423 , 623-656, October, 1948.

Sharifi B., anthony Hutton M., Kalita J., « Automatic Summarization of Twitter Topics », *in National Workshop on Design and Analysis of Algorithm*, 2010a.

Sharifi B., Hutton M.-A., Kalita J., « Summarizing Microblogs Automatically », *Human Language Technologies : The Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 685-688, 2010b.

Sharifi B., Hutton M.-A., Kalita J. K., « Experiments in Microblog Summarization », *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10, IEEE Computer Society, Washington, DC, USA, p. 49-56, 2010c.

Shou L., Wang Z., Chen K., Chen G., « Sumblr : Continuous Summarization of Evolving Tweet Streams », *the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, ACM, New York, NY, USA, p. 533-542, 2013.

Xu W., Grishman R., Meyers A., Ritter A., « A Preliminary Study of Tweet Summarization using Information Extraction », *Proceedings of the Workshop on Language Analysis in Social Media*, Association for Computational Linguistics, Atlanta, Georgia, p. 20-29, June, 2013.

Zubiaga A., Spina D., Amigó E., Gonzalo J., « Towards Real-time Summarization of Scheduled Events from Twitter Streams », *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, HT '12, ACM, New York, NY, USA, p. 319-320, 2012.