LIA@INEX2012 : Combinaison de thèmes latents pour la contextualisation de tweets

M. Morchid*, R. Dufour*, G. Linarès*

*339 chemin des Meinajaries,
84911 Avignon cedex 9
{mohamed.morchid, richard.dufour, georges.linares}@univ-avignon.fr,
http://lia.univ-avignon.fr

Résumé. La quantité d'information échangée sur Internet ne cesse de croître et prend de plus en plus souvent la forme de message courts (*tweet*, messagerie instantanée, ...). De part le peu d'informations véhiculées dans ces types de messages, il est nécessaire de connaître leur contexte d'apparition afin de les rendre compréhensibles par un lecteur. Nous présentons dans ce papier une méthode de contextualisation de messages courts utilisant une représentation thématique. Cette représentation permet d'étendre le vocabulaire du message par un ensemble de mots thématiquement proches. Cette méthode a été appliquée avec succès à la problématique de la contextualisation de *tweets* dans le cadre de la campagne d'évaluation INEX 2012 (CLEF 2012).
Les résultats obtenus montrent l'apport de cette méthode pour une meilleure compréhension de messages courts.

1 Introduction

L'augmentation exponentielle des données disponibles dans le Web permet aux utilisateurs d'accéder à une importante quantité d'information. Cependant, l'exploitation de ces données nécessite la mise en place de systèmes de recherche d'information performants, que ce soit en termes de rapidité ou de pertinence. À cette masse de données s'ajoute l'expansion rapide des plates-formes de *micro-blogging*. Ces espaces d'échanges particuliers permettent aux utilisateurs de transmettre des idées, opinions ou des faits communs sous la forme de messages courts. Selon la plate-forme d'échange utilisée, la taille de ces messages peut même être limitée à un nombre maximum de mots ou de caractères [1]. Cette contrainte liée à la taille du message entraîne l'utilisation d'un vocabulaire particulier, qui se trouve être souvent peu standard, mal orthographié ou même tronqué (Choudhury et al. (2007)), l'objectif étant d'échanger un maximum d'informations en un minimum de caractères.

Pour ces raisons, la tâche de *Question-Réponse* (QR) connaît un engouement au sein de la communauté scientifique. Le nombre de campagnes d'évaluation ne cesse de croître depuis l'organisation de la première campagne TREC [2] (*Text Retrieval Conference*) en 1999. Son objectif est de comparer les performances de différents systèmes devant répondre à des questions

1. Par exemple, la plate-forme *Twitter* n'autorise pas l'envoi de messages dont la taille dépasse 140 caractères.
2. http ://trec.nist.gov

factuelles, tous les participants devant traiter exactement le même jeu de données. Dans la continuité, la campagne CLEF [3] (*Conference and Labs of Evaluation Forums*) a été organisée en 2009 et 2010 autour de ces mêmes problématiques de QR. Cette tâche est depuis 2011 attribuée à INEX (SanJuan et al. (2011)). Les participants devaient répondre à une question sur le contexte d'un *tweet* écrit en anglais (*"what is this tweet about ?"*) en n'ayant simplement à leur disposition qu'un corpus issu de *Wikipedia* (avril 2011).

Le principe de cette tâche a changé pour INEX 2012 (SanJuan et al. (2012b)). À présent, les participants doivent utiliser un système de Recherche d'Information (RI) et un système de Résumé Automatique (RA) pour la recherche d'un contexte sachant un message court. Ce contexte est composé au plus de 500 mots issus d'articles *Wikipedia* et a pour vocation de permettre au lecteur une meilleure compréhension du *tweet*. Cette tâche peut être divisée en deux sous-tâches. La première consiste à rechercher les documents *Wikipedia* les plus pertinents en appliquant un système de RI (Schiffman et al. (2007); Pakray et al. (2010, 2011)). La seconde sous-tâche consiste à extraire les passages les plus représentatifs du *tweet* au moyen de documents *Wikipedia* pertinents. Pour cette tâche, nous disposons de l'outil *Indri* (Strohman et al. (2005b)) permettant l'indexation de phrases contenues dans des documents XML. Une table d'indexation est tout d'abord construite à partir de l'ensemble des phrases contenues dans un corpus de documents. L'outil permet ensuite d'extraire et d'ordonner les phrases de cette table selon leur proximité à une requête fournie sous format *Indri*. Cet outil devrait nous permettre de construire le contexte, à partir des documents *Wikipedia*, d'un *tweet* (assimilable ici à une requête). Il convient donc d'élaborer la requête la plus représentative du *tweet* considéré, l'ensemble des mots contenus dans un *tweet* n'étant pas forcément la requête optimale. Pour ce faire, nous disposons uniquement du contenu lexical du tweet (moins de 140 caractères). Ce vocabulaire réduit et peu standard ne permet pas de dégager aisément un ensemble de mots-clefs caractéristiques de l'idée véhiculée par le message court.

Notre proposons de contextualiser un *tweet* au moyen d'une analyse latente de Dirichlet (LDA) (Blei et al. (2003)) afin d'obtenir une représentation de ce *tweet* dans un espace thématique. Cette représentation permet de trouver un ensemble de thèmes latents composant le *tweet* ; de ces thèmes, est extrait un ensemble de mots-clefs caractéristiques. La tâche de contextualisation de *tweets* proposée par INEX 2012 permet d'éprouver le système d'extraction de mots-clefs utilisant un espace thématique que nous avons développé. L'avantage principal du système que nous proposons est son application directe à différentes tâches (extraction de mots-clefs, classification de documents, ...) sans modification, aucun des paramètres du système ne nécessitant une quelconque adaptation.

D'autres approches basées sur des modèles statistiques existent, telles que LSI/LSA (Dumais (1994); Bellegarda (1997)) ou pLSA (Hofmann (1999)). Ces méthodes ont démontré leur efficacité dans des tâches variées. Dans (Bellegarda (2000)), les auteurs proposent d'utiliser le modèle LSA (*Latent Semantic Analysis*) pour extraire les phrases les plus pertinentes d'un document audio transcrit automatiquement. Dans (Suzuki et al. (1998)), les auteurs appliquent la méthode LSA pour l'extraction de mots-clefs depuis une base de données encyclopédique.

L'identification des thèmes principaux du message permet la recherche des mots-clefs dans une représentation plus riche que son simple contenu lexical, grâce à l'analyse de grands corpus. C'est une forme d'expansion du *tweet* qui doit permettre d'améliorer la caractérisation du

3. http ://www.clef-initiative.eu

message. Ceci est particulièrement important lorsque le message est écrit dans un langage peu standard, situation assez fréquente sur la plate-forme de micro-blogging *Twitter*.

Ces mots-clefs et le *tweet* composent alors la requête Indri soumise à l'outil d'indexation *Indri*[4]. Celui-ci fournit en retour un résumé issu d'articles *Wikipedia* liées à la requête et qui sont supposé permettre de contextualiser le *tweet*. L'intérêt porté au contexte d'un tweet est une manière nouvelle d'analyser le contenu des messages de *Twitter*, ou des messages courts plus généralement. Différents autres aspects de *Twitter* ont fait l'objet d'études récentes soit dans un cas général sur son fonctionnement (Yang et al. (2010)), soit comme un espace compact fortement réactif dans lequel un ensemble de descripteurs d'opinions sont extraits (Larceneux (2007)).

Dans la prochaine partie de ce papier, nous décrirons les données utilisées par notre système. Puis la méthode proposée sera décrite dans la partie 3. La partie 4 sera consacrée aux différentes expériences menées et à l'évaluation de notre système, avant de conclure dans la partie 5.

2 COMPOSITION DES CORPUS

Pour constituer un modèle LDA robuste, une quantité importante de données est nécessaire. Dans cette optique, un corpus D de documents a été extrait à partir d'articles *Wikipedia* récents en anglais (novembre 2011). Ce corpus, fourni aux participants de la campagne d'évaluation INEX 2012 SanJuan et al. (2012b), est composé d'environ 3,7 millions d'articles. De ce corpus de documents, l'ensemble des notes et références bibliographiques ont été retirées. Chacun des documents est fourni au format XML et respecte les définitions de types de documents (DTD) décrites dans le tableau 1. Au final, ce corpus correspond à environ 26 millions de phrases pour un total d'environ 333 millions d'occurrences de mots. Le vocabulaire contient, quant à lui, 2,8 millions de mots uniques (présents au moins une fois dans le corpus).

```
<!ELEMENT xml (page)+>
<!ELEMENT page (ID, title, a, s*)>
<!ELEMENT ID (#PCDATA)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT a (p+)>
<!ELEMENT s (h, p+)>
<!ATTLIST s o CDATA #REQUIRED>
<!ELEMENT h (#PCDATA)>
<!ELEMENT p (#PCDATA | t)*>
<!ATTLIST p o CDATA #REQUIRED>
<!ELEMENT t (#PCDATA)>
<!ATTLIST t e CDATA #IMPLIED>
```

TAB. 1 – *DTD des pages Wikipedia.*

4. Indri est un moteur de recherche issu du projet *Lemur*, un travail réalisé en collaboration entre l'université du Massachusetts et l'université de Carnegie Mellon. Voir : http ://www.lemurproject.org/indri/

Le corpus de test de la campagne INEX 2012 a également été utilisé afin de vérifier la performance de notre système. Le corpus contient 1 142 *tweets* extraits à partir de *Twitter*, soit 16 263 occurrences de mots, pour un vocabulaire de 5 287 mots uniques. Chaque *tweet* est composé d'un identifiant (Id) et de son contenu textuel (un exemple est disponible au tableau 2), et n'excède pas 140 caractères.

Le contexte associé automatiquement par notre système à chaque *tweet* doit contenir au maximum 500 mots. Celui-ci est réalisé par un système de recherche d'information couplé à un système de résumé automatique fournis par les organisateurs (SanJuan et al. (2012a)). Ceux-ci regroupent :

1. Un index *Indri* recouvrant tous les mots (sans l'utilisation de liste d'arrêt ou *stemming*) et tous les tags XML.
2. Un système de *PartOfSpeech* (POS) réalisé par *TreeTagger*[5].
3. Un algorithme performant de résumé automatique créé par *TermWatch*[6] (Chen et al. (2010)).
4. L'évaluation des résumés est basée sur FRESA (Saggion et al. (2010)).

Ce système reçoit en entrée une requête dans le langage *Indri* (Metzler et Croft (2004)) puis retourne un résumé. Celui-ci est composé de phrases POS annotées avec *TreeTagger*. Ce processus d'annotation permet d'attribuer un score à chacune des phrases en utilisant *TermWatch*. Cet ensemble de phrases constitue le contexte du *tweet*.

Id	Texte
169939776420577280	celtics blog welcome to the garden celtics

TAB. 2 – *id et texte contenus dans un tweet du corpus de test INEX 2012.*

3 SYSTÈME DE CONTEXTUALISATION DE TWEETS

Le système de contextualisation de *tweets* peut être décomposé en deux sous-tâches. La première consiste à élaborer une requête à partir d'un *tweet*. La seconde sous-tâche s'intéresse à l'envoi de cette requête afin de recevoir en retour un ensemble de phrases considérées comme le contexte du *tweet*.

Concrètement, la méthode proposée enchaîne cinq étapes :

1. Estimation *off-line* d'un modèle LDA depuis un large corpus de documents D ; cette étape produit un espace thématique T_{spc} de taille $n^{T_{spc}}$ de vocabulaire $v^{T_{spc}}$ et un vecteur V^w représentant la distribution des thèmes pour chacun des mots w de $v^{T_{spc}}$; chacune des caractéristiques V_i^w est la probabilité du mot w sachant la classe z_i issue de la LDA.
2. Utilisation du *Gibbs sampling* pour inférer une distribution des classes LDA pour un tweet t avec T_{spc}. Un vecteur de caractéristiques V^t est alors obtenu ; chacune des caractéristiques V_i^t est la probabilité de la classe z_i issue de la LDA sachant le tweet t (chacune des classes peut être considérée comme un thème).

5. http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
6. http ://data.termwatch.es

3. Projection du vecteur V^t dans l'espace de vocabulaire $v^{T_{spc}}$ pour obtenir un score $s(w)$ représentant la popularité du mot w dans le tweet. Ensuite, un sous-vocabulaire S^w est composé des mots issus d'une LDA ayant obtenu les meilleurs scores $s(w)$.
4. Création de la requête q avec les mots issus du tweet t et du sous-vocabulaire S^w.
5. Envoi de la requête q à l'index *Indri* de phrases *Wikipedia* afin d'extraire un ensemble de phrases représentant le contexte c du tweet t.

Les différentes étapes de notre système de contextualisation de *tweets* sont décrites dans la figure 1. Un exemple de contextualisation d'un *tweet* au moyen de notre méthode est proposé dans la figure 2.

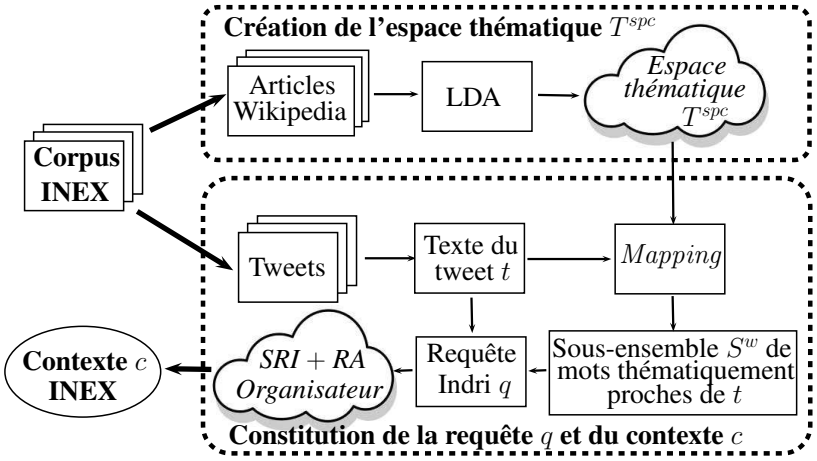

FIG. 1 – *Architecture du système de contextualisation de tweets.*

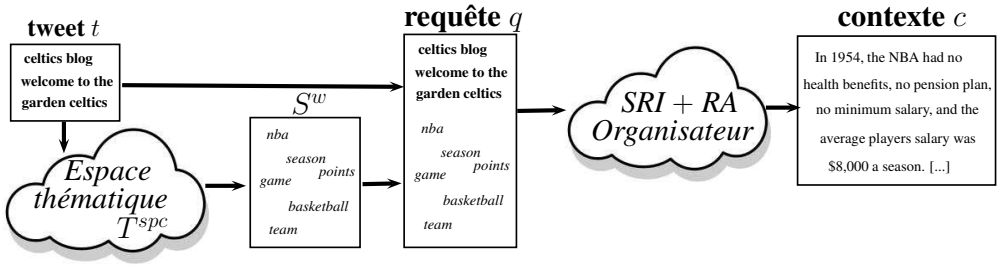

FIG. 2 – *Exemple de contextualisation d'un tweet du corpus INEX 2012.*

Les sections suivantes sont consacrées à la description détaillée des différentes étapes de contextualisation.

3.1 Vecteur de caractéristiques V^t

Le langage utilisé dans les messages *Twitter* est peu standard et ne peut contenir qu'au maximum 140 caractères. Pour ces raisons, un espace thématique T^{spc} issu d'une LDA permet d'enrichir le vocabulaire initial du *tweet*. Cette expansion du vocabulaire dans un espace de plus grande taille permet ainsi la constitution d'une requête plus robuste au simple contenu lexical du *tweet*. Un vecteur de caractéristiques V^t est ensuite constitué, chacune de ces caractéristiques permettant de représenter l'importance du thème sachant le *tweet*.

3.1.1 Espace thématique T_{spc}

L'analyse latente de Dirichlet (LDA) est un modèle génératif probabiliste qui considère le document vu comme un *sac de mots* (Salton (1989)), comme un mélange probabiliste de thèmes latents. Contrairement à un modèle de mélange de multinomiales, LDA considère qu'un thème est associé à chaque occurrence de mots composant le document, plutôt que d'associer un thème au document complet. Ainsi, un document peut changer de thèmes d'un mot à un autre. Il est toutefois à noter que les occurrences de mots sont liées par une variable latente qui contrôle le respect global de la distribution des thèmes dans le document. Ces thèmes latents sont caractérisés par une distribution de probabilités de mots qui leur sont associés. À l'issue de cette analyse LDA, un espace thématique de n_{spc} thèmes est obtenu avec pour chacun des thèmes z, la probabilité de chaque mot du vocabulaire v^{spc} sachant le thème z.

Le formalisme LDA est décrit dans la figure 3. Pour chaque document N du corpus D, un premier paramètre θ est tiré suivant une loi de Dirichlet du paramètre α. Puis un second paramètre ϕ est tiré suivant la même loi de Dirichlet sur le paramètre β. Puis pour générer chacun des mots w du document N, on tire un thème latent z depuis une distribution multinomiale sur θ. Sachant ce thème z, la distribution des mots est une multinomiale de paramètres ϕ. Le paramètre θ est tiré pour tous les documents depuis un même paramètre a priori α. Ceci permet d'avoir un paramètre liant tous les documents (Blei et al. (2003)).

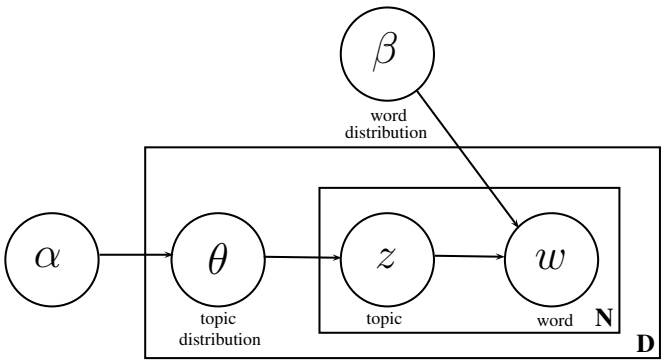

FIG. 3 – *Formalisme du modèle LDA.*

Un espace T^{spc} de 400 thèmes est obtenu par une LDA sur le corpus D. Les 30 mots les plus représentatifs du *tweet* issus du vocabulaire thématique sont sélectionnés ($|S^w| = 30$).

3.1.2 Projection de *tweets* dans l'espace thématique et élaboration du vecteur V^t

L'algorithme de *Gibbs sampling* (Griffiths et Steyvers (2002)) a été utilisé pour inférer un tweet t ainsi que l'espace de thèmes T^{spc}. Cet algorithme repose sur la méthode *Markov Chain Monte Carlo* (MCMC). Le *Gibbs sampling* permet donc d'obtenir des échantillons des paramètres de distribution θ sachant un mot du document de test w et un thème donné z_i. Un vecteur de caractéristiques V^t est alors obtenu. La i^{eme} caractéristique V_i^t (où $i = 1, 2, \ldots, n^{T_{spc}}$) est la probabilité du thème z_i sachant le tweet t :

$$V_i^t = P(z_i|t) \ . \tag{1}$$

3.2 Sous-vocabulaire S^w issu du vocabulaire $v^{T_{spc}}$

Cette méthode permet une extraction des mots les plus proches thématiquement d'un *tweet*. Un score de pertinence est donné à chacun des mots du vocabulaire $v^{T_{spc}}$. Un sous-vocabulaire S^w est constitué des 30 mots ayant le score de pertinence s le plus élevé. Le score s du mot w est la probabilité *a priori* que le mot w soit généré par le tweet t :

$$\begin{aligned} s(w) &= P(w|t) \\ &= \sum_{i=1}^{n^{T_{spc}}} P(w|z_i)P(z_i|t) \\ &= \sum_{i=1}^{n^{T_{spc}}} V_i^w \times V_i^t \\ &= \left\langle V^w, V^t \right\rangle \end{aligned}$$

où $P(w|z_i)$ est la probabilité que le mot w (où $w \in v^{T_{spc}}$) soit généré par le thème z_i. Le score s est normalisé pour être compris entre 0 (mot non pertinent) et 1 (très pertinent).

Au travers des exemples proposés dans le tableau 3, nous constatons que les mots contenus dans un *tweet* n'apparaissent pas nécessairement dans le sous-vocabulaire S^w des mots thématiquement proches. Ces exemples illustrent bien notre motivation initiale, à savoir trouver un ensemble de mots décrivant le *tweet* mais n'apparaissant pas dans celui-ci. L'approche proposée permet donc d'enrichir le vocabulaire associé à un *tweet*. Par exemple, nous pouvons constater dans le tweet (2), que certains termes génériques pour décrire l'événement (*army, war, muslim* ou *islamic*) n'apparaissent pas dans le *tweet*.

3.3 Requête *Indri* q

Nous avons choisi de composer la requête q en unifiant les mots contenus dans le tweet t avec le sous-vocabulaire S^w (voir partie 3.2). La figure 4 montre les différents éléments qui composent la requête q représentant le tweet t. Cette requête est l'association des mots contenus dans le tweet t et des 30 mots thématiquement les plus proches $S^w = \{w_1, w_2, \ldots, w_{|S^w|}\}$.

Tweet	**10 premiers mots de S^w ($\|S^w\| = 30$)**
celtics blog welcome to the garden celtics (1)	**nba** season game team points **basketball** games time year played
syrian troops attack residential areas in hama and homs (2)	**battle army** street forces troop troops **wa** **muslim** men **islamic** city
bras for after breast implant surgery 3 tips (3)	blood **heart** surgery **pain** body pressure patient patients muscle **tissue**
did you know that 2012 is the international year of sustainable energy for all you canfind out more at our (4)	development international **world environmental** **global** public human national **policy** **government**
wow childhood abuse disrupts brain formation study (5)	children **disorder mental** child **therapy** **syndrome treatment** disorders people symptoms

TAB. 3 – *Exemples de tweets avec les 10 mots ayant le score s le plus élevé. En* **gras** *quelques mots intéressants ne figurant pas dans le tweet.*

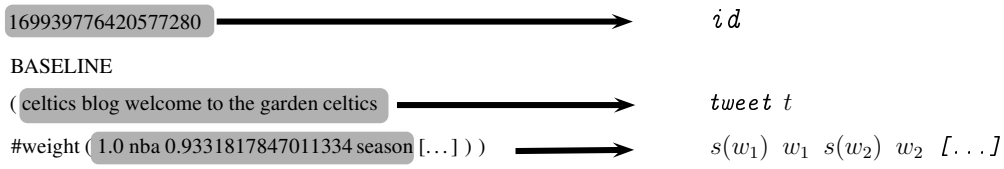

FIG. 4 – *Exemple de requête q pour un tweet t du corpus INEX 2012*

3.4 Contexte c

Cette requête q est ensuite envoyée au système fourni par les organisateur utilisant *Indri* (Strohman et al. (2005a)) pour l'indexation de paragraphes *Wikipedia* anglais afin d'obtenir un ensemble de phrases répondant au mieux à cette requête. Cet ensemble de moins de 500 mots composera le contexte c du tweet t. Le système de résumé automatique utilisant un index *Indri* est accessible par des requêtes via une interface CGI en utilisant un script perl [7].

Exemple d'un tweet t et de son contexte c :

tweet t : celtics blog welcome to the garden celtics.
contexte c : In later life, Cousy was Commissioner of the American Soccer League from 1974 to 1979, and he has been a color analyst on Celtics telecasts since the 1980s. Today, he is a marketing consultant for the Celtics, and occasionally makes broadcast appearances with Mike Gorman and ex-Celtic teammate Tom Heinsohn. In 1954, the NBA had no health benefits, no pension plan, no minimum salary, and the average players salary was $8,000 a season. [...] 147

7. http ://qa.termwatch.es/data

Boston Celtics season was the 1st season of the Boston Celtics in the Basketball Association of America (BAA/ NBA).

4 EXPÉRIMENTATION ET RÉSULTATS

Nous détaillons dans cette partie les résultats obtenus par notre système sur la problématique de la contextualisation de *tweets* selon différentes métriques. L'évaluation s'est déroulée dans le cadre de la participation à la campagne d'évaluation INEX 2012 regroupant 33 participants.

4.1 Évaluation sur le contenu informatif du contexte

L'objectif de cette métrique est d'évaluer la sélection de passages pertinents (SanJuan et al. (2011)). Dans ce cas précis, un ensemble de 63 *tweets* forment le corpus d'évaluation. Les 60 meilleurs passages [8] pour chacun des *tweets* sont sélectionnés pour l'évaluation. Ce choix est réalisé en fonction du score attribué par le système automatique de contextualisation de *tweets* (scores les plus élevés).

La dissimilarité entre un texte de référence et le résumé proposé est donnée par :

$$Dis(T,S) = \sum_{t \in T} (P-1) \times \left(1 - \frac{min(log(P), log(Q))}{max(log(P), log(Q))}\right)$$
$$P = \frac{f_T(t)}{f_T} + 1$$
$$Q = \frac{f_S(t)}{f_S} + 1$$

T représente l'ensemble des termes contenus dans le texte de référence. Pour chacun des termes $t \in T$, $f_T(t)$ représente la fréquence d'apparition de t dans le texte de référence et $f_S(t)$ sa fréquence d'apparition dans le résumé proposé à l'évaluation (SanJuan et al. (2011)). Plus $Dis(T,S)$ est faible, plus le résumé proposé est similaire au texte de référence. T peut prendre trois formes distinctes :

- Uni-gramme : un lemme unique (forme canonique du terme).
- Bi-gramme : deux lemmes successifs dans la même phrase.
- Bi-gramme *2-gaps* : de même que le bi-gramme, mais peut être séparé par deux autres lemmes.

Les résultats de notre système (*run* 193) ainsi que ceux obtenus par le système *baseline* (*run* 194, fourni par les organisateurs) et celui ayant obtenu le meilleur score (*run* 178), sont donnés dans la table 4.

4.2 Évaluation sur la facilité de lecture du contexte

Cette métrique nécessite la collaboration des participants pour évaluer l'ensemble des contextes attribués automatiquement aux 63 *tweets*. Rappelons que chacun des contextes ne

8. Le terme "passage" correspond aux phrases en sortie de l'outil *Indri*.

Run Id	Description du Run	Rang	**Métrique d'information**		
		(*sur 33*)	*Uni-gram*	*Bi-gram*	*Skip-gram*
193	Espace de thèmes	7	0.7909	0.8920	0.8938
178	Meilleur Run	1	0.7734	0.8616	0.8623
194	Baseline Organisateur	4	0.7864	0.8868	0.8887

TAB. 4 – *Résultats officiels pour la tâche de contextualisation de tweets INEX-2012 pour le contenu informatif du contexte.*

peut excéder 500 mots (SanJuan et al. (2011)). Pour chacun des passages à évaluer, le participant doit juger si le passage contient :

- *Syntaxe* (S) : une erreur de syntaxe dans le passage.
- *Anaphore* (A) : des répétitions d'un élément antérieur.
- *Redondance* (S) : une information redondante.
- *Corbeille* (T) : aucun lien avec le passage antérieur.

Le tableau 5 présente les résultats de notre système (*run* 193) ainsi que ceux obtenus par le système *baseline* (*run* 194) et celui ayant obtenu le meilleur score (*run* 185).

Run Id	Description du Run	Rang	**Métrique de lisibilité**		
		(*sur 33*)	*Pertinence*	*Syntaxe*	*Structure*
193	Espace de thèmes	12	0.6208	0.6115	0.5145
185	Meilleur Run	1	0.7728	0.7452	0.6446
194	Baseline Organisateur	4	0.6975	0.6342	0.5703

TAB. 5 – *Résultats officiels pour la tâche de contextualisation de tweets INEX-2012 pour la lisibilité du contexte.*

4.3 Évaluation non-officielle sur la précision du contexte

Chaque contexte est constitué d'un titre d'article *Wikipedia*. Cette métrique permet de mesurer la similarité entre les titres des textes de référence et les résumés à évaluer. Les résultats obtenus sont fortement corrélés avec les résultats de l'évaluation sur le contenu informatif du contexte (table 4). Trois méthodes classiques ont été choisies pour l'évaluation : la précision (mesure du bruit), le rappel (mesure du silence) et la F-mesure (moyenne arithmétique entre la précision et le rappel). Les résultats de notre système (*run* 193) ainsi que ceux obtenus par le système *baseline* (*run* 194) et celui ayant obtenu le meilleur score (*run* 152), sont donnés dans la table 6.

5 DISCUSSIONS ET CONCLUSIONS

Nous constatons que la méthode proposée obtient de bons résultats lors de l'évaluation en contenu informatif (table 4). De plus, sur les 2 146 mots issus de l'espace thématique utilisés pour la constitution de la requête Indri, 1 174 n'apparaissent pas dans le vocabulaire des

Run Id	Description du Run	Rang	**Métrique de précision**		
		(*sur 33*)	*Précision*	*Rappel*	*F-mesure*
193	Espace de thèmes	10	0.156219	0.442979	0.198238
152	Meilleur Run	1	0.321815	0.455337	0.323508
194	Baseline Organisateur	8	0.153116	0.462193	0.210242

TAB. 6 – *Résultats non-officiels pour la tâche de contextualisation de tweets INEX-2012.*

tweets (54%). Ce constat montre bien l'apport d'un vocabulaire tournant autour des thématiques proches du *tweet*. Ces ensembles de mots sont souvent absents du *tweet* et permettent alors une généralisation de l'idée véhiculée par le tweet comme cela a été détaillé dans le tableau 3. Le fait d'avoir choisi de retenir les thématiques proches du *tweet*, avec une pondération qui dépend de l'importance du thème sachant le tweet et de l'importance de chaque mot sachant le thème, a pour conséquence de privilégier des termes fortement corrélés thématiquement. Par exemple, un thème très proche d'un tweet (probabilité $P(z_i|t)$ élevée) , permettra au vocabulaire le décrivant de bénéficier de cette pondération très forte. La requête q qui en résultera, contiendra majoritairement des termes proches de ce thème. Les résultats obtenus, en terme de facilité de lecture du contexte (table 5), tiennent compte des redondances et autres anaphores. Ils peuvent être alors influencés par ce vocabulaire thématiquement très proche. Le système obtient des résultats assez comparables pour la pertinence des titres *Wikipedia* retournés (table 6) par rapport à ceux obtenus dans l'évaluation du contenu informatif du contexte.

Dans ce papier, nous avons décrit une méthode de contextualisation de *tweets* basée sur une représentation thématique. Selon la métrique considérée, notre système se classe entre la 7^{eme} et la 12^{eme} place sur les 33 systèmes proposés. La performance de notre système montre que cette approche permet une bonne contextualisation d'un message court. Cette tâche est rendue d'autant plus ardue que les messages issus de *Twitter* utilisent un vocabulaire peu standard.

Les résultats obtenus permettent d'entrevoir de nouvelles possibilités et perspectives. Celles-ci peuvent se concentrer sur plusieurs points, à savoir le choix de la pondération entre thèmes et mots pour l'attribution d'un score pour un mot du vocabulaire thématique, ou encore la modification des caractéristiques de l'index en enlevant notamment les mots outils. Il serait également intéressant d'étudier le comportement de notre système en remplaçant les mots par leurs lemmes, ou en modifiant les caractéristiques de l'espace thématique (nombre de thèmes composant l'espace, choix d'un corpus autre que *Wikipedia* ...).

6 REMERCIEMENTS

Ce travail a été réalisé dans le cadre du projet SuMACC de l'Agence National de Recherche (ANR) en vertu du contrat ANR-10-CORD-007.

Références

Bellegarda, J. (1997). A latent semantic analysis framework for large-span language modeling. In *Fifth European Conference on Speech Communication and Technology*.

Bellegarda, J. (2000). Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE 88*(8), 1279–1296.

Blei, D., A. Ng, et M. Jordan (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research 3*, 993–1022.

Chen, C., F. Ibekwe-SanJuan, et J. Hou (2010). The structure and dynamics of cocitation clusters : A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology 61*(7), 1386–1409.

Choudhury, M., R. Saraf, V. Jain, S. Sarkar, et A. Basu (2007). Investigation and modeling of the structure of texting language. In *IJCAI-Workshop on Analytics for Noisy Unstructured Text Data*, pp. 63–70.

Dumais, S. (1994). Latent semantic indexing (lsi) and trec-2. *NIST SPECIAL PUBLICATION SP*, 105–105.

Griffiths, T. et M. Steyvers (2002). A probabilistic approach to semantic representation. In *Proceedings of the 24th annual conference of the cognitive science society*, pp. 381–386. Citeseer.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI ' 99*, pp. 21. Citeseer.

Larceneux, F. (2007). Buzz et recommandations sur internet : quels effets sur le box-office ? *Recherche et applications en marketing*, 45–64.

Metzler, D. et W. Croft (2004). Combining the language model and inference network approaches to retrieval. *Information processing & management 40*(5), 735–750.

Pakray, P., P. Bhaskar, S. Banerjee, B. Pal, S. Bandyopadhyay, et A. Gelbukh (2011). A hybrid question answering system based on information retrieval and answer validation. In *CLEF 2011 Workshop on QA4MRE*.

Pakray, P., P. Bhaskar, S. Pal, D. Das, S. Bandyopadhyay, et A. Gelbukh (2010). Ju_cse_te : System description qa@ clef 2010–respubliqa. In *CLEF 2010 Workshop on Multiple Language Question Answering (MLQA 2010)*.

Saggion, H., J. Torres-Moreno, I. Cunha, et E. SanJuan (2010). Multilingual summarization evaluation without human models. In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, pp. 1059–1067. Association for Computational Linguistics.

Salton, G. (1989). Automatic text processing : the transformation. *Analysis and Retrieval of Information by Computer*.

SanJuan, E., P. Bellot, V. Moriceau, et X. Tannier (2011). Overview of the inex 2010 question answering track (qa@ inex). *Comparative Evaluation of Focused Retrieval*, 269–281.

SanJuan, E., V. Moriceau, X. Tannier, P. Bellot, et J. Mothe (2012a). Overview of the inex 2011 question answering track (qa@inex). In S. Geva, J. Kamps, et R. Schenkel (Eds.), *Focused Retrieval of Content and Structure*, Volume 7424 of *Lecture Notes in Computer Science*, pp. 188–206. Springer Berlin Heidelberg.

SanJuan, E., V. Moriceau, X. Tannier, P. Bellot, et J. Mothe (2012b). Overview of the inex 2012 tweet contextualization track. In *Copyright cG2012 remains with the author/owner (s). The unreviewed pre-proceedings are collections of work submitted before the December*

workshops. They are not peer reviewed, are not quality controlled, and contain known errors in content and editing. The proceedings, published after the Workshop, is the authoritative reference for the work done at INEX., pp. 148.

Schiffman, B., K. McKeown, R. Grishman, et J. Allan (2007). Question answering using integrated information retrieval and information extraction. In *Proceedings of NAACL HLT*, pp. 532–539.

Strohman, T., D. Metzler, H. Turtle, et W. Croft (2005a). Indri : A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*.

Strohman, T., D. Metzler, H. Turtle, et W. B. Croft (2005b). Indri : A language model-based search engine for complex queries. In *International Conference on Intelligence Analysis*.

Suzuki, Y., F. Fukumoto, et Y. Sekiguchi (1998). Keyword extraction using term-domain interdependence for dictation of radio news. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pp. 1272–1276. Association for Computational Linguistics.

Yang, Z., J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, et Z. Su (2010). Understanding retweeting behaviors in social networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 1633–1636. ACM.

Summary

The amount of information exchanged over the Internet is growing and becoming more and more as short messages (tweet, HMI, ...). Due to the limited information conveyed in these types of messages, it is necessary to know its context to make them understandable by users. In this paper, we present a method of contextualization of short messages using a thematic representation. This representation allows to extend the vocabulary of short messages by a set of thematically related words. This method has been successfully applied to the problem of tweet contextualization in the context of INEX2012 evaluation benchmark (CLEF2012).
The results show the contribution of this method to a better understanding of short messages.