Atelier Contextualisation de Messages Courts

Organisateurs : Patrice Bellot (LSIS / Polytech Marseille), Josiane Mothe (IRIT), Eric SanJuan (LIA), Ludovic Tanguy (CLLE-ERSS)

PRÉFACE

Cet atelier s'intéresse à la contextualisation de messages courts. Ces messages courts peuvent correspondre à des requêtes soumises à un moteur de recherche, un post sur twitter, un SMS,...
La contextualisation a pour objectif d'aider le lecteur à comprendre le texte court (définition de termes, principales thématiques visualisation d'éléments externes en lien avec le texte, ...) ou de permettre à un système automatique de le traiter au mieux (expansion de requêtes, résumé, typage). Dans tous les cas, elle consiste en un apport de connaissance à un support textuel minimal.
La contextualisation peut ainsi être déclinée selon des aspects liés à l'utilisateur (profil, historique de production ou d'interaction, localisation...) ou à l'environnement du message (liens entre documents, type de document...). Les méthodes et techniques utilisées pour cette contextualisation pourront faire appel à des travaux multidisciplinaires: linguistique, mathématiques, informatique...

Pour cet atelier, nous avons retenu quatre présentations :

- Les sessions de recherche comme contexte des requêtes, Simon Leva
- LIA@INEX2012 : Combinaison de thèmes latents pour la contextualisation de tweets, Mohamed Morchid, Richard Dufour, Georges Linarès
- Création de snippets : une application de la génération automatique de résumés, Liana Ermakova, Nicolas Faessel
- Contextualisation de messages courts : l?importance des métadonnées, Jean-Valère Cossu, Julien Gaillard, Juan-Manuel Torres-Moreno, Marc El-Bèze

Des tables rondes sur le thème de la création de collections de référence complèteront l'atelier.

Patrice Bellot	Josiane Mothe	Eric SanJuan	Ludovic Tanguy
LSIS / Polytech Marseille	IRIT	LIA	CLLE-ERSS

Membres du comité de lecture

Le Comité de Lecture est constitué de:

Patrice Bellot, Université de Marseille
Jean-Pierre Chevallet, Université de Grenoble
Cécile Fabre, Université de Toulouse
Véronique Moriceau, Université de Paris-Sud 11
Josiane Mothe, Université de Toulouse
Eric San-Juan, Université d'Avignon
Jacques Savoy, Université de Neuchâtel, Suisse
Xavier Tannier, Université de Paris-Sud
Ludovic Tanguy, Université de Toulouse

TABLE DES MATIÈRES

Les sessions de recherche comme contexte des requêtes

Simon Leva*

*CLLE-ERSS : CNRS et Université de Toulouse (UMR 5263),
5 allées Antonio Machado, 31058 Toulouse Cedex 9
sleva@univ-tlse2.fr
http://w3.erss.univ-tlse2.fr/textes/pagespersos/leva/

Résumé. La tâche d'identification des sessions des utilisateurs d'un moteur de recherche a suscité la construction de plusieurs collections de référence et l'élaboration de multiples méthodes de détection automatique. Cette tâche constitue en effet le point de départ de nombreuses études s'intéressant au contexte de la recherche et aux besoins d'information des utilisateurs. Nous détaillons dans cette étude la construction d'une collection de référence à partir d'un journal de requêtes issu du portail *OpenEdition*, et nous présentons une évaluation des annotations manuelles constituant cette collection. La référence obtenue contient 947 requêtes pour 406 sessions, avec un taux d'accord (Kappa de Cohen) entre les annotateurs allant de $0,47$ à $0,61$. Cette collection servira à l'évaluation de méthodes de détection automatique des sessions ainsi qu'à des études portant sur les reformulations de requêtes.

1 Introduction

Les utilisateurs d'un moteur de recherche sur le Web font appel à différentes stratégies afin de satisfaire leurs besoins d'information. En particulier, les requêtes soumises peuvent faire l'objet de plusieurs reformulations visant à préciser le besoin d'information initial. Les requêtes d'un utilisateur font ainsi souvent partie d'une session de recherche, et ne devraient pas être considérées de manière isolée. En effet, une session fournit de nombreux indices sur le contexte de la recherche, l'objectif de l'utilisateur ou son expertise dans le domaine considéré. Les reformulations de requêtes et les documents consultés sont des éléments particulièrement utiles pour une meilleure compréhension du besoin d'information, de son évolution et de sa satisfaction au fil d'une recherche. La notion de session est donc une notion clef en recherche d'information, et plus spécifiquement en recherche d'information contextuelle.

L'étude présentée dans cet article porte sur la construction d'une collection de référence manuellement annotée au niveau des sessions, ainsi que sur l'évaluation de la concordance des annotations. La collection constituée provient d'un journal de requêtes issu du portail de ressources électroniques dédiée aux sciences sociales *OpenEdition* [1]. La constitution d'une collection de référence dans cet environnement vise deux objectifs principaux : 1) servir de référence pour l'évaluation de méthodes de détection automatique des sessions ; 2) servir de

1. http ://www.openedition.org/

référence pour des études sur les types de reformulations effectuées par les utilisateurs de la plateforme, et à terme pour réaliser une typologie des requêtes.

Ce travail s'inscrit dans le cadre du projet ANR CAAS *(Contextual Analysis and Adaptive Search)* — programme Contint — coordonné par Josiane Mothe (IRIT), et faisant l'objet d'un partenariat entre l'Institut de Recherche en Informatique de Toulouse (IRIT), le Laboratoire Informatique d'Avignon (LIA) et l'Équipe de Recherche en Syntaxe et Sémantique du laboratoire Cognition, Langue, Langage, Ergonomie (CLLE-ERSS).

Dans cet article, nous nous intéressons tout d'abord à la notion de session telle que définie en recherche d'information, ainsi qu'aux collections de référence précédemment constituées (section 2). Puis, nous présentons la méthodologie de construction de la collection de référence issue du portail *OpenEdition* (section 3). Enfin, nous détaillons la collection construite et l'évaluation des annotations (section 4), avant d'envisager les perspectives de ce travail.

2 La notion de session en recherche d'information

2.1 Définitions de la notion de session

L'une des premières définitions de la notion de session dans le cadre de l'étude d'un journal de requêtes a été proposée par Silverstein et al. (1999) :

> *A* session *is a series of queries by a single user made within a small range of time. A session is meant to capture a single user's attempt to fill a single information need.*

Cette idée de regroupement des activités d'un utilisateur (soumission d'une requête, navigation sur la page de résultats, consultation d'un résultat) correspondant à un thème spécifique se retrouve chez Göker et He (2000), qui insistent également sur la proximité temporelle entre les activités au sein d'une session :

> *This paper focuses on the temporal ordering of activities clustered according to close proximity in time. [...] We group activities and refer to the resulting unit as a session. If we view a user with an interest in a specific topic as acting in a particular role, then [...] the activities in the same session are likely to correspond to one role.*

Spink et al. (2006) proposent une définition moins restrictive de la notion de session, qui est considérée comme une simple séquence de requêtes d'un utilisateur :

> *A session is the entire series of queries submitted by a user during one interaction with the Web search engine. Session length varied from less than a minute to a few hours.*

Cette définition est ensuite affinée par Jansen et al. (2007) et correspond alors à un épisode de recherche, notion distinguée de celle de session :

> *One can define a user episode on a Web search engine as a temporal series of interactions among a searcher, a Web system, and the content provided by that system within a specific period. [...] However, it is possible that one searching episode will be composed of one or more sessions. We define a session from a contextual viewpoint as a series of interactions by the user toward addressing a single information need.*

Tandis que les définitions précédentes envisagent les requêtes d'une session sous la forme de séquences, impliquant une certaine contiguïté entre les requêtes, Jones et Klinkner (2008) développent une vision hiérarchique de la notion de session. Ainsi, une session se compose d'une ou de plusieurs missions de recherche, qui se composent à leur tour d'un ou de plusieurs buts pouvant donner lieu à une ou plusieurs requêtes. De plus, les requêtes liées à un même but peuvent être imbriquées avec des requêtes visant un autre but :

> *A search session is all user activity within a fixed time window. [...] A search goal is an atomic information need, resulting in one or more queries. [...] The queries need not be contiguous, but may be interleaved with queries from other goals. [...] A search mission is a related set of information needs, resulting in one or more goals.*

Gayo Avello (2009) reprend la distinction entre épisode et session de recherche opérée par Jansen et al. (2007) tout en adoptant à nouveau une vision séquentielle de la notion de session :

> *[A searching episode] refers to the actions performed by a particular user within a search engine during, at most, one day. Such a searching episode can comprise one or more sessions where each of these includes one or more successive queries related to one single information need or goal.*

L'idée d'imbrication au sein d'une session entre des requêtes correspondant à des buts distincts est finalement reprise par Lucchese et al. (2011) :

> *Task-based sessions [are] sets of possibly non contiguous queries issued by the user of a Web Search Engine for carrying out a given task.*

Les définitions de la notion de session proposées dans la littérature se distinguent à plusieurs niveaux. Une session peut ainsi contenir des éléments de différente nature : des activités/interactions (Göker et He, 2000; Jansen et al., 2007) ou des requêtes (Silverstein et al., 1999; Spink et al., 2006; Jones et Klinkner, 2008; Gayo Avello, 2009; Lucchese et al., 2011). Cette différence est directement liée aux informations disponibles dans les journaux de requêtes, selon qu'ils fournissent uniquement les requêtes soumises par les utilisateurs, ou incluent également leurs actions. D'autre part, les sessions se distinguent par leur durée, plutôt courte (Silverstein et al., 1999; Göker et He, 2000) ou pouvant aller jusqu'à quelques heures (Spink et al., 2006), mais dans tous les cas inférieure à une journée (Gayo Avello, 2009). Cette contrainte temporelle s'explique notamment par le renouvellement de l'adresse IP des utilisateurs toutes les 24 heures, rendant impossible leur différenciation au-delà de cette période. Malgré ces points de divergence, la majorité des définitions s'accordent sur le fait qu'une session permet de regrouper des actions ou des requêtes liées à un même besoin d'information. Ces actions ou ces requêtes peuvent alors être envisagées de manière séquentielle (Silverstein et al., 1999; Göker et He, 2000; Jansen et al., 2007; Gayo Avello, 2009) ou imbriquée (Jones et Klinkner, 2008; Lucchese et al., 2011). Ce dernier cas est fréquemment rencontré lors d'une recherche multitâche *(multitasking search)* (Spink et al., 2006), l'utilisateur effectuant alors une recherche sur plusieurs thèmes simultanément.

2.2 Collections de référence existantes

Plusieurs initiatives de constitution de collections de référence ont vu le jour dans le cadre de la détection automatique des sessions. Les collections ainsi constituées ont contribué au

développement de méthodes de détection automatique, notamment à travers des techniques d'apprentissage, mais ont également servi à l'évaluation de ces méthodes.

Göker et He (2000) ont mené une campagne d'annotation sur un journal de requêtes issu du réseau Intranet de l'agence de presse *Reuters* pour l'année 1999. Deux annotateurs experts en formulation de requête ont été chargés d'identifier les sessions de 1 440 utilisateurs pour un ensemble de 9 534 requêtes. La collection ainsi constituée a servi de référence pour l'évaluation d'une méthode de détection automatique basée sur la définition d'un seuil temporel entre les activités faisant partie d'une même session.

Jansen et al. (2007) ont procédé à l'annotation d'un échantillon de 2 000 requêtes soumises au méta-moteur de recherche *Dogpile* durant l'année 2005. Ces annotations ont été comparées à trois méthodes de détection automatique des sessions exploitant les informations d'adresse IP et de *cookies* soit seules, soit en les associant respectivement à un seuil temporel et à des patrons de reformulation de requêtes. L'exploitation de ces annotations a notamment permis d'identifier l'origine des erreurs générées par les méthodes automatiques évaluées.

Jones et Klinkner (2008) ont réalisé l'annotation d'un échantillon de 3 jours de requêtes soumises au moteur de recherche *Yahoo !* au cours de l'année 2007. La collection de référence constituée se compose de 2 922 sessions pour un ensemble de 8 226 requêtes soumises par 312 utilisateurs. La consigne donnée pour l'annotation était que les requêtes appartenant à une même session possèdent les mêmes critères de réussite en termes de satisfaction du besoin d'information de l'utilisateur. Cette collection leur a permis de tester l'opérabilité de leur définition de session, mission et but de recherche. L'annotation réalisée est de type ascendant, le groupe d'annotateurs ayant exploité les pages de résultats et les clics des utilisateurs afin de déterminer leurs buts, puis leurs missions et finalement leurs sessions. De plus, le cas des requêtes liées à un même but et imbriquées au sein d'une même session est envisagé.

Gayo Avello (2009) a mis en place une campagne d'annotation portant sur des échantillons de sept journaux de requêtes provenant de moteurs de recherche commerciaux (*AlltheWeb*, *Altavista*, *AOL* et *Excite*). La collection ainsi constituée comporte près de 95 000 requêtes soumises par près de 15 000 utilisateurs, pour un ensemble de près de 35 000 sessions manuellement identifiées. Il s'agit de l'initiative la plus ambitieuse de constitution d'une collection de référence. Un annotateur expert a été chargé d'évaluer si deux requêtes successives sont ou non thématiquement reliées. L'annotateur s'est basé uniquement sur le texte des requêtes, mais pouvait éventuellement faire appel à une ressource externe en cas de manque de connaissances. Cette collection de référence a servi de cadre commun pour l'évaluation et la comparaison de diverses méthodes de détection automatique.

Nous constatons que les différentes annotations manuelles des sessions recensées dans la littérature contiennent relativement peu de détails concernant le protocole d'annotation en lui-même. En particulier, aucune précision n'est indiquée quant aux difficultés rencontrées par les annotateurs. Il n'existe ainsi à notre connaissance aucune étude proposant d'évaluer l'accord inter-annotateur pour la tâche de détection des sessions. Par ailleurs, les études mentionnées n'indiquent pas de référence permettant d'accéder aux collections constituées.

3 Méthodologie de construction d'une référence annotée

3.1 Définition d'une session

Nous retenons dans le cadre de ce travail les définitions suivantes :
- un *épisode de recherche* correspond à l'ensemble des requêtes soumises à un moteur de recherche par un utilisateur donné durant au plus une journée ; cet épisode de recherche peut contenir une ou plusieurs sessions de recherche ;
- une session de recherche correspond à l'ensemble des requêtes reliées à un même besoin d'information ; ces requêtes peuvent être imbriquées au sein d'un même épisode de recherche dans le cas d'un épisode multitâche.

3.2 Annotation manuelle des sessions

La tâche d'annotation des sessions consiste à indiquer, pour un utilisateur et un épisode de recherche donné, à quelle session appartient chacune des requêtes incluses dans cet épisode. Autrement dit, il s'agit de segmenter les épisodes de recherche de chaque utilisateur en une ou plusieurs sessions. Nous avons élaboré un guide d'annotation pour familiariser les annotateurs avec les notions mobilisées tout en détaillant les aspects pratiques de la tâche.

Une session = un ensemble de requêtes liées Pour chaque épisode de recherche d'un utilisateur, les annotateurs doivent observer l'ensemble des requêtes soumises avant de prendre une décision et d'attribuer une session à chaque requête. Ce point est crucial, car le lien entre deux requêtes peut ne pas être évident de prime abord, mais apparaître dans une requête ultérieure. Par exemple, le lien n'est pas évident entre la requête `barriere de corail` suivie de la requête `nouvelle zélande`. Pourtant, ce lien s'éclaire avec la soumission d'une troisième requête `barriere de corail nouvelle zélande`. Il ne s'agit donc pas de considérer uniquement les couples de requêtes successives.

Des requêtes liées à un même besoin d'information Si les requêtes regroupées au sein d'une session visent un même besoin d'information, celui-ci n'apparaît pas toujours de manière explicite. Afin d'aiguiller les annotateurs, nous leur avons proposé une liste d'indices potentiels, pouvant apparaître de manière combinée. Ces indices relèvent de différents niveaux d'information :

1. *Indices textuels :* les requêtes possèdent des mots en commun, ou des parties de mot en commun. C'est par exemple le cas des requêtes `femmes moralistes`, `femmes moralistes 18 siecle` et `moralistes`.
2. *Indices sémantiques :* les requêtes possèdent des mots qui ne sont pas strictement identiques au niveau orthographique, mais qui sont néanmoins liés par une relation sémantique telle que la synonymie, l'hyperonymie ou l'hyponymie. Par exemple, les requêtes `foresterie` et `sylviculture` sont reliées par le fait que le terme *foresterie* est un synonyme en français québécois pour le terme *sylviculture*.
3. *Indices de proximité thématique :* les requêtes possèdent des mots qui ne sont ni identiques, ni liés par une relation sémantique classique, mais qui restent néanmoins lexicalement proches dans le cadre de thématiques ou d'objets spécifiques. Par exemple,

les requêtes `théâtre expérimental` et `grotowski` sont reliées par le fait que le polonais Jerzy Grotowski était un metteur en scène et théoricien du théâtre.

Utilisation de ressources externes En l'absence d'indices textuels, il peut parfois être nécessaire de faire appel à une ressource externe (dictionnaire, thésaurus, encyclopédie, etc.) afin de palier un manque de connaissances. Les annotateurs doivent alors indiquer la ressource utilisée et expliciter l'élément ayant permis de lever l'indécision. C'est par exemple le cas pour les requêtes `ifriqya` et `tunisie`, une réponse possible étant d'indiquer que l'encyclopédie *Wikipédia* mentionne *Le territoire de l'Ifriqiya correspond aujourd'hui à la Tunisie*.

Une session possède un identifiant numérique unique au sein d'un épisode Afin d'identifier chacune des sessions d'un utilisateur, il est demandé aux annotateurs d'attribuer à chaque requête d'un épisode le numéro de la session à laquelle elle appartient. La première session de chaque épisode est identifiée par 1, la seconde par 2, et ainsi de suite. Ce type de numérotation permet également d'identifier les requêtes imbriquées reliées à une même session.

4 Collection de référence

La collection de référence que nous avons constituée est issue de données provenant du portail *OpenEdition*. Nous présentons tout d'abord cet environnement de recherche avant de détailler le journal de requêtes exploité et les résultats de l'annotation de la collection.

4.1 Le portail *OpenEdition*

Le portail *OpenEdition* propose un libre accès à un ensemble de ressources électroniques dans le domaine des sciences humaines et sociales. Développé et dirigé par le Centre pour l'édition électronique ouverte (Cléo), il se compose de trois plateformes dont chacune est dédiée à une ressource électronique spécifique : *Revues.org* diffuse 353 revues et 22 collections de livres, *Calenda* recense plus de 20 000 évènements scientifiques en lettres et en sciences humaines et sociales, tandis qu'*Hypotheses.org* héberge 269 blogs et carnets de recherche.

Plusieurs points d'entrée permettent d'effectuer une recherche dans cet environnement varié. D'une part, un moteur de recherche principal est accessible sur la page d'accueil du portail *OpenEdition* et de la plateforme *Revues.org*. D'autre part, une recherche peut également se faire directement à partir du moteur de recherche situé sur le site d'une revue associée à *Revues.org*. Dans ces deux situations, les résultats sont présentés dans une interface commune, permettant de préciser le type d'information recherchée à l'aide de champs (titre, auteur, résumé, etc.) et de restreindre la recherche à l'aide de filtres (plateforme de publication et type de document visé, année de la publication, etc.).

4.2 Journal de requêtes exploité

Nos travaux s'appuient sur un journal de requêtes provenant du portail *OpenEdition*. Ce journal de requêtes contient une collection de 1 057 471 requêtes soumises par 227 302 utilisateurs durant la période du 07 avril 2010 au 1er février 2012. Les requêtes sont majoritairement formulées en français, mais certaines sont également en anglais ou en espagnol. À la différence

des requêtes soumises à un moteur de recherche généraliste, la particularité de l'environnement d'*OpenEdition* fait que les requêtes émanent principalement d'acteurs du monde académique, et ciblent des revues, des évènements ou des blogs.

La construction de ce journal de requêtes a nécessité la mise en œuvre de plusieurs traitements afin de ne conserver que les informations les plus fiables à partir du journal d'accès original. En particulier, les données ont été nettoyées et filtrées de manière à éliminer les informations inexploitables (requêtes soumises par des robots d'indexation, longues séquences de requêtes strictement identiques, suites de signes de ponctuation, etc.) et à contenir des requêtes provenant d'utilisateurs individuels. Les requêtes ont ensuite été regroupées par adresse IP et classées par ordre chronologique. Le journal de requêtes finalement obtenu comporte un identifiant pour chaque utilisateur — correspondant à l'adresse IP anonymisée —, la date et l'heure de soumission de chaque requête, ainsi que les requêtes soumises.

La tâche d'annotation des sessions a été menée sur un échantillon du journal de requêtes *OpenEdition* suffisamment large pour être représentatif des phénomènes en présence tout en restant aisé à traiter pour les annotateurs. Nous avons ainsi sélectionné aléatoirement une collection de 947 requêtes soumises par 216 utilisateurs. Cet échantillon a été automatiquement segmenté en 349 épisodes de recherche, correspondant pour chaque utilisateur à l'ensemble de ses requêtes soumises en une journée au plus [2]. Trois annotateurs — dont l'auteur de cette étude —, non spécialistes dans les domaines représentés par les documents, ont été chargés de segmenter les requêtes de chaque épisode de recherche en une ou plusieurs sessions.

Utilisateur	Requête	Épisode	Session
39	travail saisonnier	1	1
	industries de loisirs		2
	parc d'attraction		2
	fidélisation et emplois saisonniers		1
	travail saisonnier		1
	parc à thèmes		2
	parc astérix		2
	disneyland		2
	disneyland paris		2
	walibi		2
	compagnies des alpes		2
26	génération	1	1
	les jeunes		1
	ruralité		2

TAB. 1 – *Épisodes annotés en sessions extraits de la collection OpenEdition.*

Le tableau 1 présente un extrait de la collection *OpenEdition* après annotation. Pour chaque utilisateur est précisé l'ensemble de ses requêtes ainsi que les épisodes correspondants. Nous indiquons ici l'annotation de chaque épisode en sessions réalisée par l'annotateur 1.

2. Dans le cas d'une séquence de requêtes soumises avant et après minuit, chevauchant donc deux journées, la mise en place d'un seuil temporel évite la séparaison en deux épisodes distincts.

Le cas de l'utilisateur 39 correspond à un épisode multitâche, illustrant l'imbrication entre des requêtes relevant de thématiques de recherche différentes. Nous pouvons effectivement identifier d'une part un besoin d'information lié au thème du travail saisonnier et d'autre part un besoin d'information lié au thème des parcs de loisirs. Ces besoins d'information ont donné lieu à deux sessions distinctes, respectivement marquées par les identifiants 1 et 2, et ce malgré une discontinuité entre les requêtes.

Le cas de l'utilisateur 26 constitue un exemple de difficulté d'annotation. Si le lien est clair entre les deux premières requêtes, la troisième pose problème : s'agit-il d'une précision de la thématique précédente, auquel cas la requête porterait sur la ruralité selon les générations, ou s'agit-il d'une nouvelle thématique ? Étant donné qu'il n'y a ici aucune indication explicite de lien entre les requêtes, par exemple sous la forme d'une quatrième requête `ruralité chez les jeunes`, l'annotateur 1 a considéré qu'il s'agit de deux thématiques et donc de deux sessions distinctes.

4.3 Résultats de l'annotation

	Annotateur 1	Annotateur 2	Annotateur 3
Nombre de sessions	393	428	410

TAB. 2 – *Nombre de sessions identifiées dans la collection par chaque annotateur.*

Le tableau 2 présente le nombre de sessions identifiées dans la collection *OpenEdition* par chaque annotateur. Si ce nombre varie, il reste cependant proche de 410 sessions identifiées en moyenne. Afin de comprendre les différences au niveau du nombre de sessions identifiées et d'estimer le taux de concordance entre les annotateurs, nous confrontons chaque paire d'annotations obtenues dans une matrice de confusion. Ce type de représentation permet de visualiser, pour un nombre de catégories identique entre deux annotateurs, le nombre d'annotations communes et d'annotations différentes entre ces annotateurs.

		Ann. 2			Ann. 3		
		CS	NS	Total	CS	NS	Total
Ann. 1	CS	**490**	57	547	**514**	33	547
	NS	12	**39**	51	11	**40**	51
	Total	502	96	**598**	525	73	**598**

TAB. 3 – *Matrices de confusion des annotations effectuées par l'annotateur 1 et les annotateurs 2 et 3.*

Les tableaux 3 et 4 représentent les matrices de confusion des annotations obtenues pour chaque paire d'annotateurs. Nous n'exploitons pas directement les identifiants de sessions attribués à chaque requête : dans les cas de détection de plus de deux sessions et d'épisode multitâche, une nouvelle session peut en effet recevoir un identifiant différent entre les annotateurs

		Ann. 3		
		CS	NS	Total
Ann. 2	CS	**482**	20	502
	NS	43	**53**	96
	Total	525	73	**598**

TAB. 4 – *Matrice de confusion des annotations effectuées par les annotateurs 2 et 3.*

même si ces derniers s'accordent sur le fait qu'une requête marque le début d'une nouvelle session et non la continuation de la session précédente. Une telle situation est représentée dans le tableau 5, où la requête `citoyen définition` fait partie de la session 2 pour l'annotateur 1 et de la session 3 pour l'annotateur 3, tout en constituant le début d'une nouvelle session pour les deux annotateurs. Afin d'éviter ce biais, nous nous basons donc sur l'identification par les annotateurs d'une nouvelle session (noté NS dans les matrices de confusion et le tableau 5) ou d'une continuation de la session précédente (noté CS dans les matrices de confusion et le tableau 5) au sein d'un même épisode. De plus, nous ne prenons pas en compte lors de la constitution des matrices de confusion les cas triviaux — un épisode de recherche ne contient qu'une seule requête et constitue donc une seule session — pour lesquels aucune alternative ne s'offre aux annotateurs, aboutissant à une annotation identique. Cela explique le fait que seules 598 requêtes sont comptabilisées au total au lieu des 947 initialement annotées.

Utilisateur	Requête	Épisode	Ann. 1		Ann. 3	
142	`Flandre Wallonie`	2	1	NS	1	NS
	`BHV`		1	CS	2	NS
	`conflit périphérie`		1	CS	1	NS
	`conflit périphérie Bruxelles`		1	CS	1	CS
	`citoyen définition`		**2**	**NS**	**3**	**NS**
	`définition citoyen`		2	CS	3	CS

TAB. 5 – *Exemple de désaccord au niveau des identifiants de session et d'accord au niveau de la détection d'une nouvelle session entre les annotateurs 1 et 3.*

Par exemple, nous voyons dans le tableau 3 qu'il y a 39 requêtes pour lesquelles l'annotateur 1 et l'annotateur 2 ont considéré qu'elles marquent le début d'une nouvelle session au sein d'un même épisode, constituant donc des annotations identiques, tandis qu'il y a 12 requêtes pour lesquelles l'annotateur 1 a considéré qu'elles marquent le début d'une nouvelle session tandis que l'annotateur 2 les a annotées comme une continuation de la session précédente.

Nous utilisons le coefficient Kappa (Cohen, 1960) afin d'évaluer le degré de concordance entre chaque paire d'annotateurs au niveau de l'identification d'une requête comme débutant une nouvelle session ou poursuivant la session précédente. Ce coefficient se base sur une différence relative entre l'accord réel observé et un accord aléatoire, rendant ainsi possible une

comparaison. Il se définit par le rapport :

$$\kappa = \frac{P_a - P_e}{1 - P_e}$$

où P_a est la proportion d'accord observée entre deux annotateurs, et P_e la proportion d'accord aléatoire. Le coefficient Kappa correspond donc à une estimation de l'accord optimal entre les annotateurs après avoir retranché la part de cet accord dû au hasard — un accord parfait correspondant à un κ égal à 1. Le tableau 6 présente le coefficient Kappa obtenu pour chaque paire d'annotateurs. Le taux d'accord varie de modéré ($0,47$ et $0,57$) à bon ($0,61$).

Paires d'annotateurs	κ
Annotateurs 1 et 2	$0,47$
Annotateurs 1 et 3	$0,61$
Annotateurs 2 et 3	$0,57$

TAB. 6 – *Accord inter-annotateur estimé avec le coefficient Kappa.*

Utilisateur	Requête	Épisode	Session ann. 1	Session ann. 2
7	`bank`	1	1	1
	`bank crisis`		1	1
	`China party`		2	1
	`china financial`		2	1
	`bank`		1	1
72	`philosophie`	1	1	1
	`jean lasrière`		1	2
	`jean ladrière`		1	2
	`"jean ladrière"`		1	2

TAB. 7 – *Exemples de désaccord entre les annotateurs 1 et 2.*

Le tableau 7 présente des exemples de désaccord au niveau de l'identification des sessions entre les annotateurs 1 et 2. Pour l'utilisateur 7, l'annotateur 1 a considéré que les requêtes `China party` et `china financial` portent sur une thématique à part entière, tandis que l'annotateur 2 les a reliées avec la thématique des autres requêtes de cet épisode, relative à la banque. Il s'agit d'un cas difficile, car s'il est possible de trouver un lien entre ces requêtes à l'aide des termes *financial* et *bank*, il est également possible de séparer les requêtes contenant le terme *bank* de celles contenant le terme *china*. Pour l'utilisateur 72, l'annotateur 1 a identifié un lien que l'annotateur 2 n'a pas trouvé : Jean Ladrière était un philosophe et logicien belge. Nous mettons donc en avant deux types d'erreurs susceptibles d'intervenir lors de l'identification des sessions, provenant d'une part de la difficulté de choisir le lien le plus pertinent parmi plusieurs liens possibles, et d'autre part de la difficulté d'exploitation systématique d'une ressource externe pour les liens sémantiques faibles.

4.4 Collection de référence constituée

Étant donnée la présence de points de désaccord entre les annotateurs, nous avons constitué notre collection de référence en sélectionnant pour chaque requête les annotations faisant l'objet d'un accord entre au moins deux annotateurs, ou, lorsque les identifiants de session diffèrent entre les trois annotateurs, en prenant en compte la détection ou non d'une nouvelle session. Pour les 947 requêtes constituant la collection de référence, 406 sessions ont ainsi été identifiées, chaque session contenant en moyenne $2,33$ requêtes. Ce résultat est comparable au cas d'une collection provenant d'un moteur de recherche généraliste sur le Web, pour lequel chaque session contient entre $2,33$ et $2,96$ requêtes (Gayo Avello, 2009).

5 Conclusion

Dans cet article, nous avons détaillé les différentes étapes de construction d'une collection de référence annotée au niveau des sessions. La collection ainsi constituée se compose de 947 requêtes soumises par 216 utilisateurs sur le portail *OpenEdition*, correspondant à 406 sessions de recherche. Nous avons également proposé une mesure de l'accord entre les annotations obtenues en utilisant le coefficient Kappa de Cohen. Nous avons ainsi montré que la tâche d'annotation manuelle des sessions de la collection possède un taux de concordance globalement modéré entre les annotateurs, avec un Kappa allant de $0,47$ à $0,61$.

Les cas de désaccord entre les annotateurs confirment qu'il s'agit d'une tâche non triviale, et soulèvent plusieurs difficultés essentiellement liées à l'identification du besoin d'information de l'utilisateur. Il s'avère notamment difficile de trancher lorsque les requêtes contiennent peu de mots et que ces derniers possèdent un sens suffisamment large pour donner lieu à une multiplicité de liens possibles entre les requêtes.

La collection de référence constituée servira à l'évaluation de diverses méthodes de détection automatique des sessions, et notamment de méthodes traitant le cas des épisodes multitâches. Nous pourrons alors procéder à la segmentation en sessions du journal de requêtes *OpenEdition* dans son ensemble à l'aide de la méthode la plus appropriée. Cette étape constitue le point de départ d'une étude plus globale sur les reformulations de requêtes, se basant sur le fait que chaque session est marquée par un contexte de recherche caractéristique observable à travers l'ensemble des requêtes soumises. Ces requêtes pourront alors faire l'objet d'une typologie en fonction de la similarité entre les contextes qui les contiennent.

6 Remerciements

Nous adressons nos plus vifs remerciements à Marin Dacos et l'équipe du Cléo pour leur collaboration et l'accès aux données du portail *OpenEdition*. Nous tenons également à remercier nos annotateurs Clémentine et Nicolas pour leur disponibilité et l'intérêt porté à ce travail.

Références

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement 20*(1), 37–46.

Gayo Avello, D. (2009). A Survey on Session Detection Methods in Query Logs and a Proposal for Future Evaluation. *Information Sciences 179*(12), 1822–1843.

Göker, A. et D. He (2000). Analysing Web Search Logs to Determine Session Boundaries for User-Oriented Learning. In P. Brusilovsky, O. Stock, et C. Strapparava (Eds.), *Adaptive Hypermedia and Adaptive Web-Based Systems*, Volume 1892 of *Lecture Notes in Computer Science*, pp. 319–322. Springer-Verlag Berlin Heidelberg.

Jansen, B. J., A. Spink, C. Blakely, et S. Koshman (2007). Defining a Session on Web Search Engines. *Journal of the American Society for Information Science and Technology 58*(6), 862–871.

Jones, R. et K. L. Klinkner (2008). Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 699–708.

Lucchese, C., S. Orlando, R. Perego, F. Silvestri, et G. Tolomei (2011). Identifying Task-Based Sessions in Search Engine Query Logs. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pp. 277–286.

Silverstein, C., H. Marais, M. Henzinger, et M. Moricz (1999). Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum 33*(1), 6–12.

Spink, A., M. Park, B. J. Jansen, et J. Pedersen (2006). Multitasking During Web Search Sessions. *Information Processing and Management 42*(1), 264–275.

Summary

Identifying the sessions of a search engine's users aroused the creation of several benchmark collections and the elaboration of different automatic detection methods. This task actually represents the starting point of numerous studies dealing with the research context and the users' information needs. We detail within this study the creation of a benchmark collection based on a query log from the *OpenEdition* portal. We present as well an evaluation of the manual annotations which constitute this collection. The resulting benchmark contains 947 queries corresponding to 406 sessions, with an inter-annotator agreement (Cohen's Kappa) varying from 0.47 to 0.61. This collection will be exploited for both the evaluation of sessions automatic detection methods and studies of query reformulations.

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.

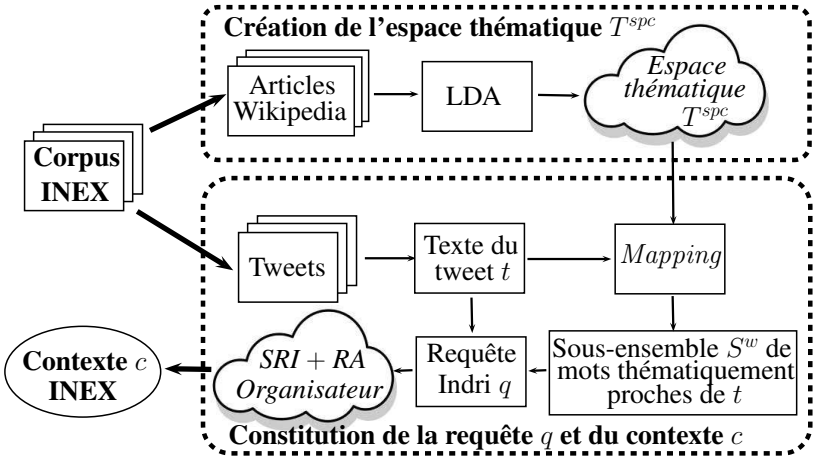

FIG. 1 – *Architecture du système de contextualisation de tweets.*

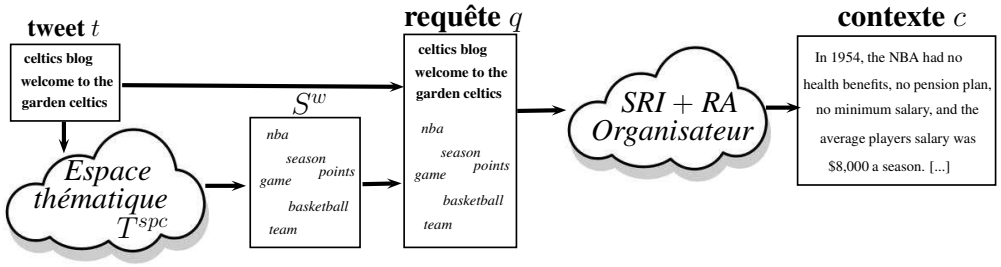

FIG. 2 – *Exemple de contextualisation d'un tweet du corpus INEX 2012.*

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)).

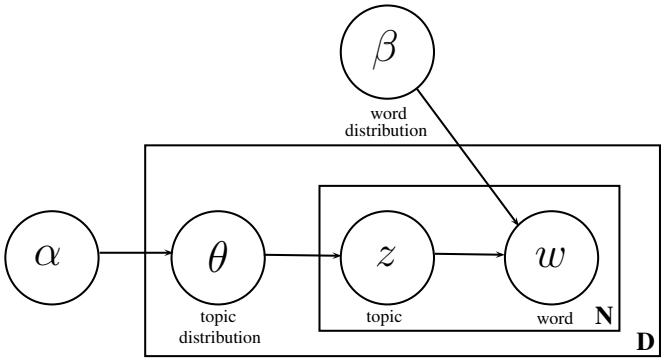

FIG. 3 – *Formalisme du modèle LDA.*

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.*

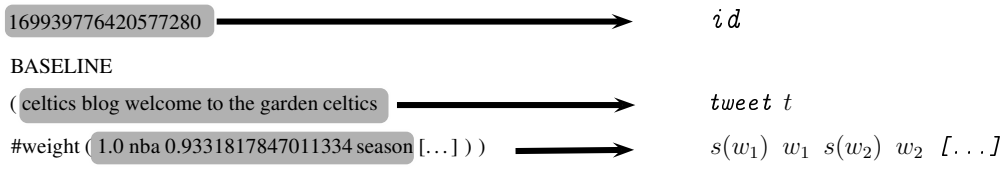

FIG. 4 – *Exemple de requête q pour un tweet t du corpus INEX 2012*

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.

Création de snippets : une application de la génération automatique de résumés

Liana Ermakova*, Nicolas Faessel**

*IRIT - UMR 5505, Université de Toulouse, France,
Université Nationale de Recherche d'État de Perm, Russie
liana.ermakova@irit.fr
**IRIT - UMR 5505, Université de Toulouse, France
nicolas.faessel@irit.fr

Résumé. Face à l'immense volume de documents renvoyé par les moteurs de recherche sur le Web, les utilisateurs sont aidés dans leur tâche de sélection de documents par de courts extraits d'une à deux phrases associés à l'URL de chaque document renvoyé par le moteur de recherche, nommés *snippets*. Dans cet article, nous considérons la génération automatique de snippets comme une tâche de résumé automatique de documents, dans le contexte d'un requête donnée. Selon notre approche, les phrases des documents susceptibles de résumer un document par rapport à une requête sont représentées par différents vecteurs prenant en compte leur contexte local et leurs caractéristiques linguistiques. Afin d'assurer la qualité de l'information fournie par les snippets en respectant la contrainte de faible longueur, nous proposons deux algorithmes pour la sélection des passages candidats. Cette approche a été testée sur la collection de la tâche de génération de snippets d'INEX 2011.

1 Introduction

Les moteurs de recherche sur le Web donnent souvent comme résultat un grand nombre de documents qu'il est impossible de lire dans leur totalité. Afin d'aider l'utilisateur à déterminer si une page Web est pertinente pour une requête sans avoir à cliquer sur le lien de cette page, les moteurs de recherche fournissent un *snippet*. Un snippet est un extrait textuel du document, provenant de son contenu ou de ses méta-données (Turpin et al., 2007) et apparaissant sous le lien pointant vers une page résultant d'une recherche. Idéalement, un snippet fournit à l'utilisateur l'information qu'il recherche. De bons snippets devraient se baser sur de petites unités d'information limitées en taille telles qu'une phrase ou encore un élément XML ou (X)HTML et ils devraient permettre de distinguer les documents qu'ils représentent des autres résultats fournis par le moteur de recherche (Huang et al., 2008).

Dans cet article, nous considérons la génération de snippets comme une tâche de résumé automatique mono-document. En effet, les snippets sont des extraits courts, généralement de moins de 300 caractères, qui doivent représenter le contenu du document, indépendamment ou non d'une requête d'utilisateur. Notre approche permet une génération contextuelle de snippets, qui dépend directement des requêtes de utilisateurs : pour une requête donnée, il s'agit

de fournir à l'utilisateur un snippet qui lui permet de savoir si le document peut répondre à son besoin d'information. Nous générons les snippets à partir de phrases ou de passages du document, en proposant une représentation multi-vectorielle prenant en compte : (1) le besoin d'information de l'utilisateur exprimé par une requête (contexte global) (2) les caractéristiques linguistiques du contenu du document, (3) la structure du document, (4) le contexte local des passages candidats à la génération de snippets, qui correspond à l'influence de leurs phrases voisines.

Compte tenu du fait que les snippets doivent être le plus informatif possible malgré leur petite taille, nous proposons, dans le cas de la sélection de passages, d'utiliser deux algorithmes : une approche dynamique pour résoudre le « problème du sac à dos » (Kellerer et al., 2004) et un algorithme de fenêtre glissante (MV). Le problème du sac à dos consiste à remplir un sac dont la capacité en terme de poids est limitée, avec des éléments qui ont une valeur et un poids. Le but est de maximiser la valeur des éléments du sac en respectant la limite de poids. Ce problème peut être appliqué dans notre contexte comme suit : sélectionner le sous-ensemble de phrases candidates dont la similarité à une requête donnée est maximale et le poids total (qui correspond ici au nombre de caractères des phrases) est inférieur ou égal à un seuil prédéfini. Toutefois, l'information la plus pertinente peut se trouver dans une phrase qui dépasse à elle seule le seuil. Pour résoudre ce problème, nous proposons un algorithme de fenêtre glissante, qui permet de rechercher des passages candidats dans des phrases. Ainsi, si les phrases qui contiennent une information pertinente sont trop grandes, l'algorithme de fenêtre glissante va permettre la sélection d'un passage pertinent, sans tenir compte de sa position de départ dans la phrase, et donc réduire le passage candidat pour la génération de snippets. Le problème de cette approche est qu'elle peut abaisser le niveau de lisibilité du snippet. Notre but est donc de combiner la capacité informative du snippet avec sa lisibilité.

L'article est organisé comme suit : la section 2 présente un état de l'art sur les approches de résumé automatique et la génération de snippets. Puis, dans la section 3 nous décrivons notre approche pour la recherche de passages candidats et la génération de snippets. Cette approche est évaluée dans la section 4 sur la collection d'INEX 2011. Enfin la section 5 conclut cet article.

2 Travaux connexes

Un snippet provenant du moteur de recherche Google est défini comme suit : « Un snippet est la description ou un extrait d'une page Web, qui suit le titre et précède l'URL et le lien Cache » (Spencer, 2010).

Certains moteurs de recherche fournissent des informations détaillées pour certaines requêtes spécifiques (snippets enrichis). Par exemple, Google gère des snippets enrichis basés sur les Microdate, Microformats et les RDFa pour les types de contenu suivants : personnes, produits, organisations, recettes ...

Les descriptions contenues dans les méta-données sont couramment utilisées pour la génération de snippets. Toutefois, les moteurs de recherche peuvent pénaliser les descriptions de mauvaise qualité (ayant par exemple une mauvaise mise en forme, trop de mots clés, des informations redondantes dans la description, le titre et le contenu, etc.) (Spencer, 2010). Un moteur de recherche peut aussi générer un snippet, non pas à partir du contenu de la page, mais à partir de la description qu'il en a (Yahoo Dictory or DMOZ) (Slawski, 2009). De plus,

Yahoo se base aussi bien sur une pertinence dépendante des requêtes que sur la capacité d'un passage à résumer le document dont il provient, indépendamment de toute requête (Kanungo et Metzler, 2009). Cette capacité est estimée à partir de caractéristiques indépendantes de la requête telles que la position du passage dans le document, le nombre de termes communs entre le passage et le titre du document, etc. Lorsque la pertinence d'un passage est dite dépendante d'une requête, elle correspond bien souvent à la similarité entre la requête et le passage (généralement calculée à partir du nombre de termes que le passage et la requête ont en commun). En plus de leurs fréquences, la distance entre les termes semble être une caractéristique importante (Wang et al., 2012). Les techniques d'expansion de requête traditionnelles comme le retour de pertinence (Leal et al., 2012),(Wang et al., 2012), (Ko et al., 2008) ou l'analyse du contexte local (Sanderson, 1998) sont utilisées pour la génération de snippets.

La lisibilité est une des propriétés clés des snippets qui doit être prise en compte par les moteurs de recherche. Kanungo et Orr (2009) suggèrent de prédire la lisibilité des snippets en appliquant des arbres de décision augmentés par le gradient (gradient boosting decison trees) pour un ensemble de caractéristiques telles que la longueur moyenne des mots, la quantité des fragments de snippets, etc. Leur étude sur le comportement des utilisateurs à travers leur clics ont montré que la lisibilité influence l'utilisateur. Clarke et al. (2007) supposent que de simples caractéristiques telles que la présence de tous les termes de la requête, la lisibilité du snippet et la longueur de l'URL peuvent significativement influencer la façon dont vont cliquer les utilisateurs. Il existe deux façons d'améliorer la lisibilité des snippets : le filtrage et la pénalisation. Le filtrage signifie que les snippets candidats qui sont illisibles doivent être filtrés, alors que la pénalisation implique que le candidat doit être pénalisé dû à sa faible lisibilité. Contrairement aux tâches de génération de résumés ou de contextualisation, l'ordonnancement des phrases n'impacte pas la lisibilité des snippets, ces derniers étant trop courts.

3 Description de la méthode

Notre méthode de génération de snippets est une adaptation de la méthode proposée par Ermakova et Mothe (2012) pour la contextualisation de tweets. Nous avons modifié les techniques de pondération et développé différents algorithmes pour l'extraction des phrases candidates à la génération de snippets.

3.1 Représentation multi-vectorielle de phrases

Un document est représenté par un ensemble de phrases. Nous modélisons une phrase comme un ensemble de vecteurs. Le premier vecteur représente les termes qui apparaissent dans une phrase. Cela correspond à une représentation unigramme. Le second vecteur correspond aux bigrammes. Les travaux de Ermakova et Mothe (2012) montrent que la comparaison d'entités nommées est efficace pour contextualiser des tweets portant sur des articles de journaux. Ainsi le troisième vecteur est composé des entités nommées identifiées dans la phrase. Au lieu de stocker les fréquences d'occurrence des composants dans chaque vecteur, nous nous contentons de stocker seulement les composants, considérant qu'un terme, un bigramme ou encore une entité nommée n'apparait rarement plus d'une fois dans une même phrase. Le fait d'effectuer une comparaison paire à paire des composants nous permet d'utiliser une représentation vectorielle creuse, c'est-à-dire ne comprenant que les composants retrouvés.

D'après Silber et Mccoy (2002), les noms fournissent l'information la plus importante. Nous proposons donc d'introduire différents coefficients permettant de faire la distinction entre l'impact des noms, des autres termes significatifs et des mots vides de sens. Ainsi, nous proposons de donner un coefficient à chaque étiquetage grammatical. Par exemple, les déterminants ont un poids nul, les noms propres ont le poids le plus fort, et les noms communs ont un poids plus fort que les verbes, adjectifs et adverbes. Les valeurs du vecteur d'unigrammes sont multipliées par le coefficient de leur valeur d'étiquetage grammatical.

La pondération des étiquettes grammaticales permet de pénaliser les anaphores pronominales non résolues et autres problèmes de lisibilité. Contrairement à la contextualisation et à la génération de résumés, nous ne pénalisons pas les phrases nominales telles que les titres, qui sont généralement très courtes et donnent une idée concise et condensée du document qui les contient. Toutefois, nous exploitons la structure des documents, en pondérant les phrases selon leur position dans les parties du document. Ainsi, les phrases provenant de résumés ont un poids plus fort que celles provenant des sections.

La sélection des phrases candidates s'effectue dans le contexte des requêtes des utilisateurs, que nous nommons ici le contexte global. Ainsi, nous générons une représentation multi-vectorielle pour chaque requête, de la même manière que pour les phrases candidates. L'appariement entre les requêtes et les phrases candidates est effectué au moyen de différentes mesures de similarité.

Pour les vecteurs unigrammes et bigrammes, nous calculons la similarité du cosinus entre une phrase et une requête (respectivement $similarity_{unigram}$ et $similarity_{bigram}$). Les vecteurs d'entités nommées sont traités différemment : pour chaque entité nommée dans une requête nous cherchons toutes les entités nommées correspondantes dans les phrases candidates. Si une requête ne contient pas d'entité nommée, toutes les phrases candidates sont considérées comme pertinentes à l'égard de ce type d'information. La similarité entre les entités nommées est notée comme suit :

$$NE_{COEF} = \frac{NE_{common} + 1}{NE_{query} + 1} \quad (1)$$

où NE_{common} est le nombre d'entités nommées apparaissant à la fois dans la phrase et dans la requête, NE_{query} est le nombre d'entités nommées de la requête. Nous lissons le résultat en ajoutant 1 au numérateur et au dénominateur : une phrase peut ne pas contenir d'entité nommée et être pourtant pertinente. Si le lissage n'était pas effectué, ce coefficient serait égal à zéro. En plus de prendre en considération les entités nommées présentes dans la phrase, nous prenons en compte les synonymes contextuels, déterminés au moyen d'un système de résolution d'anaphore et lors de la comparaison, nous choisissons le synonyme qui correspond le plus.

Cet ensemble de vecteurs nous permet de combiner les mesures de similarités obtenues pour différents types d'information. Le score final des phrases par rapport à une requête est donné par la somme pondérée de ces mesures de similarité.

3.2 Lissage en fonction du contexte local

Nous supposons que pour bien détecter une phrase candidate (appelée ici phrase cible) nous devons prendre en compte son contexte local, qui correspond aux phrases qui l'entourent.

Nous supposons que l'importance du contexte diminue au fur et à mesure que la distance augmente. Ainsi, les phrases les plus proches produisent plus d'effet sur la phrase cible que les autres phrases plus éloignées. Pour les phrases dont la distance est supérieure à k, le coefficient d'importance (c'est-à-dire le poids) est égal à zéro .

Ce système permet de prendre en compte les k phrases voisines avec un poids dépendant de leur distance de la phrase cible (équation 2). Dans ce cas, le score total R_t de la phrase cible correspond à la somme pondérée des scores des phrases voisines r_i et de la phrase cible r_0 :

$$R_t = \sum_{i=-k}^{k} w_i \times r_i \tag{2}$$

$$w_i = \begin{cases} \frac{1-w_t}{k+1} \times \frac{k-|i|}{k} & \text{si } 0 < |i| \leq k \\ w_t & \text{si } i = 0 \\ 0 & \text{si } |i| > k \end{cases} \tag{3}$$

$$\sum_{i=-k}^{k} w_i = 1 \tag{4}$$

où w_t est le poids de la phrase cible défini par l'utilisateur, w_i sont les poids des phrases du contexte k. Les poids diminuent au fur et à mesure que la distance augmente. Si la distance de la phrase dans le contexte droit ou gauche est inférieure à k, son poids est ajouté au poids w_t de la phrase cible. La contrainte exprimée dans l'équation 4 nous permet de garder la somme des poids égale à 1.

3.3 Sélection de passage

3.3.1 Problème du sac à dos

Un snippet est généralement limité à une ou deux phrases (150-300 caractères). Toutefois, il doit fournir le plus d'information possible à propos du document qu'il représente. On peut ainsi considérer la génération de snippets comme la sélection de passages d'une importance maximale et dont le poids total (c'est-à-dire la longueur) ne dépasse pas un seul prédéfini. Cela nous donne un problème classique en optimisation combinatoire : le problème du sac à dos. Ce problème est défini comme suit : étant donné un ensemble d'éléments possédant un poids et une valeur, trouver le sous-ensemble de cet ensemble permettant de remplir le sac à dos de telle façon que le poids total soit inférieur ou égal à la capacité du sac, et la valeur totale soit la plus grande possible (Kellerer et al., 2004). Nous considérons le poids comme le nombre de caractères d'une phrase, et sa similarité à la requête représente la valeur. Nous ne traitons que le problème du sac à dos 0-1 (0-1 KP), qui restreint le nombre de chaque type d'élément à zéro ou un, afin que les snippets n'aient pas de redondance d'information. Nous résolvons ce problème par la programmation de l'algorithme DP-1 avec une complexité d'exécution $o(nc)$ où n est le nombre d'éléments et c est la capacité du sac à dos (Kellerer et al., 2004).

3.3.2 Fenêtre glissante

Deux problèmes majeurs se posent lors de l'utilisation du problème du sac à dos :

1. Si chaque phrase d'un document est plus grande que le seuil prédéfini (la capacité du sac à dos), alors le snippet sera vide.
2. Cet algorithme a un temps d'exécution pseudo-polynomial.

Nous utilisons donc une fenêtre glissante pour choisir le passage ayant le meilleur score. Pour cela nous générons un nouveau passage en respectant les étapes suivantes :

1. le premier terme est enlevé du passage candidat ;
2. le terme suivant le passage candidat est ajouté tant que la taille totale du nouveau passage ne dépasse pas le seuil prédéfini ;
3. le score du nouveau passage est calculé ;
4. si le score est plus grand que le score maximal courant, il devient le nouveau score maximal.

Bien que cela permette d'améliorer le score du passage candidat, le fait que ce passage puisse commencer en milieu de phrase risque de dégrader la lisibilité. Pour éviter cela, nous proposons de pénaliser les snippets qui ne commencent pas en début de phrase.

4 Évaluation

4.1 Corpus

L'évaluation de notre système a été effectuée sur le corpus de la tâche de recherche de snippets d'INEX 2011. La collection de documents correspond à une version XML de plus de deux millions de pages provenant de Wikipédia anglais. Les expressions d'un besoin d'information sont au nombre de 50. Chaque besoin d'information contient une requête textuelle courte (titre), une requête portant sur la structure et le contenu (titre cas), une phrase de titre, une description du besoin, et une partie narrative expliquant le besoin d'information (Trappet et al., 2012).

4.2 Mesures d'évaluation

L'évaluation a été réalisée manuellement. Nous avons utilisé les mêmes techniques d'évaluation que celles utilisées dans la tâche de recherche de snippets lors de la campagne INEX 2011 (Trappet et al., 2012). Pour chaque besoin d'information, le but des évaluateurs était de déterminer si les snippets fournissaient une information suffisante concernant les documents, afin que l'utilisateur décide de la pertinence d'un document à la simple lecture du snippet correspondant. Pour cela, les évaluateurs devaient évaluer les résultats de deux manières :

– évaluation de la pertinence des documents,
– évaluation de la pertinence des snippets.

Le titre des requêtes, leur description et leur intention donnent une idée du besoin d'information de l'utilisateur. Les évaluateurs doivent parcourir les snippets et décider si le document auquel il correspond semble pertinent pour la requête seulement en lisant le snippet. La valeur de pertinence est binaire : 1 s'il semble pertinent, 0 sinon. Après cela, ils doivent lire le document entier pour juger de sa pertinence. Le jugement des documents sert de vérité terrain, permettant de comparer les jugements de pertinence des snippets.

Tout comme pour la tâche d'INEX 2011, nous avons utilisé les mesures suivantes :

– Prédiction d'exactitude moyenne (Mean accuracy prediction - MPA), qui correspond au pourcentage moyen des résultats que les évaluateurs ont correctement évalués par rapport à la vérité terrain (c'est à dire par rapport à la pertinence des documents) :

$$MPA = \frac{(TP + TN)}{(TP + FN + TN + FP)} \tag{5}$$

où TP correspond aux vrais positifs (c'est-à-dire, les snippets évalués comme pertinents et dont la vérité terrain du document correspondant l'indique comme pertinent), TN correspond aux vrais négatifs et FP et FN respectivement aux faux positifs et faux négatifs.

– Prédiction normalisée d'exactitude moyenne (MNPA), qui est le pourcentage moyen des résultats pertinents que les évaluateurs ont correctement évalués comme pertinents ainsi que les résultats non pertinents correctement évalués comme tel :

$$MNPA = 0,5 * \frac{(TP)}{(TP + FN)} + 0,5\frac{(TN)}{(TN + FP)} \tag{6}$$

– Rappel (R), qui est le pourcentage moyen des snippets pertinents évalués correctement comme tel :

$$R = \frac{TP}{(TP + FN)} \tag{7}$$

– Rappel négatif (NR), qui est le pourcentage moyen des snippets non pertinents correctement évalués comme tel :

$$NR = \frac{TN}{(TN + FP)} \tag{8}$$

– Accord positif (PA) qui est la probabilité conditionnelle d'un accord entre l'évaluateur d'un snippet et l'évaluateur d'un document, étant donné que l'un des deux a déclaré un document pertinent :

$$PA = 2 * \frac{TP}{(2 \cdot TP + FP + FN)} \tag{9}$$

– Accord négatif (NA) qui est la probabilité conditionnelle d'un accord entre l'évaluateur d'un snippet et l'évaluateur d'un document, étant donné que l'un des deux a déclaré un document non pertinent :

$$PA = 2 * \frac{TN}{(2 \cdot TN + FP + FN)} \tag{10}$$

– Moyenne géométrique (GM) du rappel et du rappel négatif :

$$GM = \sqrt{R \times NR} \tag{11}$$

4.3 Résultats

Pour chaque besoin d'information, nous avons produit une liste ordonnée de 10 documents ainsi que les snippets correspondants. Nous avons évalué deux exécutions obtenues en appliquant le problème du sac à dos (knapsack) et l'algorithme de fenêtre glissante pour la sélection

	MPA	MNPA	R	NR	PA	NA	GM
knapsack	0.81	0.81	0.76	0.86	0.80	0.83	0.81
MV	0.76	0.75	0.63	0.87	0.72	0.79	0.74

TAB. 1 – *Résultats*

de passages (MV). Les résultats présentés dans le tableau 1 montrent que l'application du problème du sac à dos donne un score bien plus élevé que la sélection par fenêtre glissante, ce malgré sa complexité de calcul.

Afin de mesurer la corrélation entre les résultats de nos deux exécutions, nous avons calculé le coefficient de contingence phi qui montre une forte corrélation entre ces variables (Everitt et Skrondal, 2010) :

$$\varphi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\cdot}n_{0\cdot}n_{\cdot 1}n_{\cdot 0}}} = 0.68 \quad (12)$$

5 Conclusion

Dans cet article, nous considérons la génération de snippets comme une tâche de résumé automatique mono-document. Nous nous appuyons sur l'approche proposée par Ermakova et Mothe (2012), qui se base sur une représentation multi-vectorielle des phrases utilisant le contexte et un ensemble de paramètres de pondération dépendant de critères aussi bien linguistiques que structurels. Par ailleurs, nous proposons deux algorithmes spécifiques à la sélection de passages candidats : une approche dynamique de résolution du problème du sac à dos, et un algorithme de fenêtre glissante. Nous avons évalué nos résultats sur les données de la tâche de recherche de snippets provenant de la campagne d'évaluation d'INEX 2011. Nos résultats montrent que l'algorithme de résolution du problème du sac à dos offre de bonnes performances malgré sa complexité.

Les perspectives de travail futur concerne la recherche des critères indépendants des requêtes et d'identification de l'intention de l'utilisateur. De plus, dans notre approche, nous avons supposé que certains paramètres utilisés dans le cas du résumé multi-document doivent être adaptés au cas de la génération de snippets. En effet, la longueur très courte des snippets laisse supposer que l'ordre des phrases n'a pas d'importance pour la génération, que les phrases nominales ne doivent pas être pénalisées, que celles provenant des titres sont utiles à la création de snippets alors qu'elles ne le sont à priori pas dans une tâche de contextualisation de tweets. Nous souhaitons valider ces suppositions dans de futures expérimentations.

Références

(2012). *10th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2011)*, Volume 7424 of *LNCS*. Springer.

Clarke, C. L. A., E. Agichtein, S. Dumais, et R. W. White (2007). The influence of caption features on clickthrough patterns in web search. In *Proceedings of the 30th annual in-*

ternational ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands, pp. 135–142. ACM.

Ermakova, L. et J. Mothe (2012). IRIT at INEX : question answering task. In pro (2012), pp. 219–227.

Everitt, B. S. et A. Skrondal (2010). *The Cambridge Dictionary of Statistics*. Cambridge University Press.

Huang, Y., Z. Liu, et Y. Chen (2008). eXtract : a snippet generation system for XML search. *VLDB 1*(2), 1392–1395.

Kanungo, T. et D. Metzler (2009). System and method for automatically ranking lines of text. Patent Application. US 2009/0292683 A1.

Kanungo, T. et D. Orr (2009). Predicting the readability of short web summaries. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, Barcelona, Spain, pp. 202–211. ACM.

Kellerer, H., U. Pferschy, et D. Pisinger (2004). *Knapsack problems*. Springer-Verlag, Berlin.

Ko, Y., H. An, et J. Seo (2008). Pseudo-relevance feedback and statistical query expansion for web snippet generation. *Inf. Process. Lett. 109*(1), 18–22.

Leal, L., F. Scholer, et J. Thom (2012). RMIT at INEX 2011 snippet retrieval track. In pro (2012).

Sanderson, M. (1998). Accurate user directed summarization from existing tools. In *Proceedings of the seventh international conference on Information and knowledge management*, Bethesda, Maryland, United States, pp. 45–51. ACM.

Silber, H. G. et K. F. Mccoy (2002). Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics 28*(4), 487–496.

Slawski, B. (2009). How a search engine may choose search snippets - SEO by the sea. http://www.seobythesea.com/2009/12/how-a-search-engine-may-choose-search-snippets/.

Spencer, S. (2010). Anatomy of a google snippet. http://searchengineland.com/anatomy-of-a-google-snippet-38357.

Trappet, M., S. Geva, A. Trotman, F. Scholer, et M. Sanderson (2012). Overview of the INEX 2011 snippet retrieval track. In pro (2012).

Turpin, A., Y. Tsegay, D. Hawking, et H. E. Williams (2007). Fast generation of result snippets in web search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, Amsterdam, The Netherlands, pp. 127–134. ACM.

Wang, S., Y. Hong, et J. Yang (2012). PKU at INEX 2011 XML snippet track. In pro (2012).

Summary

A search engine returns to a user a large volume of results associated to a query that it is impossible to read. Therefore, to define whether a web page is relevant or not to a query without clicking a link, a search engine provides a user with a small text passage appear-

ing under a search result called "snippet". In this paper, we consider snippet generation as a single-document summarization task. We propose an approach to select sentences for a snippet based on multi-vector sentence representation, taking into account the influence of neighboring sentences (smoothing from local context) and linguistic features. Snippets should be as informative as possible despite they often consist of 1-2 sentences. We propose two algorithms for the candidate passage selection: dynamic programming approach to solve the knapsack problem and the moving window algorithm. The approach was tested on the collection of INEX 2011 Snippet Retrieval Track.

Contextualisation de messages courts : l'importance des métadonnées

Jean-Valère Cossu*, Julien Gaillard*,** Juan-Manuel Torres-Moreno*,*** Marc El-Bèze*

*Laboratoire Informatique d'Avignon - Université d'Avignon et des Pays de Vaucluse
339 chemin des Meinajaries, BP91228 84911 Avignon Cedex 9, France
{jean-valere.cossu,julien.gaillard,juan-manuel.torres,marc.elbeze}@univ-avignon.fr,
http://lia.univ-avignon.fr/
**INRIA Sophia-Antipolis
julien.gaillard@inria.fr,
http://inria.fr/
***École Polytechnique de Montréal - Département de génie informatique
CP 6079 Succ. Centre Ville H3C 3A7 Montréal (Québec), Canada

Résumé. Les recherches présentées portent sur l'analyse de données fournies par le site Vodkaster. Dans l'optique d'alimenter un système de recommandation (SR) basé sur le contenu, nous developpons un système permettant, grâce à l'apprentissage automatique, de prédire la catégorie de note d'une critique. Cette critique est appelée micro-critique (μC) (140 caractères ou moins). Durant cette tâche, la prise en compte de l'identité de l'utilisateur et du titre du film, en les intégrant tels quels dans le sac de mot de la μC, a permis d'améliorer globalement les performances du système et ce, quelle que soit la méthode utilisée pour la prédiction. Nous montrons qu'avec un nombre égal de documents, l'ensemble d'apprentissage le plus proche (temporellement parlant) des données de test donnait de meilleurs résultats que les documents plus éloignés dans le temps. Ceci nous amène à envisager pour la suite de considérer une métadonnée additionnelle : la dimension temporelle.

1 Introduction

La recommandation de contenu connaît un essor important ces dernières années. Cet essor est rendu nécessaire par l'accroissement considérable du contenu disponible, conséquence de la démocratisation des nouvelles technologies de l'information. La masse de contenus est telle qu'un besoin nouveau est apparu : accéder à un contenu nouveau et différent de ce que l'on a l'habitude de consulter. L'utilisateur a besoin d'aide afin de trouver une référence qui lui convient dans des catalogues qui vont grandissant. La réponse à ce besoin peut-être fournie par un « moteur de recommandation » prenant en compte plusieurs éléments susceptibles d'aider l'utilisateur à faire le bon choix. Cependant, les moteurs de recommandation actuels se basent pour l'instant uniquement sur les notes attribuées au contenu. Un contenu est recommandé ou non selon un système complexe de statistiques ce qui peut induire plusieurs problèmes. Par exemple, le départ à froid Meyer et al. (2011) pour un contenu ou utilisateur.

Notons également leur incapacité à savoir quels sont les éléments qui les ont poussé à écarter ou mettre un avant un contenu. Un moyen de connaître ces éléments serait de prendre appui sur la critique établie par l'utilisateur, qui est accessible au système avec la note. Les évolutions en la matière des moteurs de recommandation sont soutenues par les avancées des recherches et notamment pour ce cas précis celles qui sont menées dans le domaine du Traitement Automatique de Langues (TAL). En effet, l'accès à Internet permet à un nombre toujours croissant de personnes d'exprimer leur avis sur les sites, forums de discussions et autres blogs. Les entreprises ne sont pas en reste avec les collectes d'opinions via les questionnaires électroniques, c'est ainsi que d'immenses quantités de données et d'opinions sont collectées.

C'est là que la fouille d'opinion devient utile pour extraire, de manière rapide et automatisée, les opinions des utilisateurs sur un produit ou service. Le principe devient d'autant plus intéressant avec une visée commerciale pour les grandes entreprises, disposant de masses de données de ce type (enquêtes et centres d'appel où les clients ont pu exprimer leur opinion, mais également pour les portails communautaires rassemblant les fans de cinéma, lecture, etc.). Nous proposons ici de traiter deux aspects importants de la recommandation de contenu :

– l'avis général exprimé sur un contenu : nous aborderons cette tâche comme un problème de catégorisation automatique de textes qui consiste à rattacher des textes à une ou plusieurs classes prédéfinies. Ici, les catégories correspondent à des niveaux d'opinion (positive, négative ou neutre). Le but est donc d'attribuer à chaque critique une de ces trois catégories, reflétant l'avis majoritairement exprimé par l'utilisateur.
– la recherche de chaînes caractéristiques de ces niveaux d'opinions permettant ainsi de pouvoir extraire les points les plus souvent critiqués en bien comme en mal ainsi que les marqueurs d'un certain niveau d'opinion sur ces points critiqués.

Ce traitement sera fait via une approche différente de l'accoutumée du fait de l'utilisation de la critique comme élément principal de classification et non uniquement la note. Notre travail s'articule de la manière suivante : premièrement, nous nous basons sur la répartition des critiques en fonction de leur catégorie dans le but d'extraire les termes (ou chaînes [1]) relatives à chacune des catégories d'opinions. Nous pourrons également visualiser ces chaînes pour savoir quels aspects du produit/service ont été critiqués, mais aussi les chaînes spécifiques à chaque classe. Nous reviendrons néanmoins plus en détails sur ce point en section 3. Ensuite, les chaînes extraites sont utilisées dans un système automatique de catégorisation supervisée de textes.

Le système ne considérant ainsi plus uniquement des mots isolés, mais des expressions pouvant être bien plus discriminantes.

Les méthodes permettant d'extraire les mots ou chaînes de mots et leur intérêt pour la catégorisation d'opinion sont présentées en section 2. Les techniques utilisées pour la catégorisation sont décrites en section 3. La méthode est testée sur un corpus de critiques de cinéma écrites en français (provenant d'un portail communautaire) présenté en section 4.

1. La notion de chaîne se se définit par un n-gramme agglutinant un ou plusieurs termes. Ceux-ci n'ont pas de pouvoir discriminant en étant considérés seul mais qui une fois associés à un autre terme ont un pouvoir discriminant important.

2 Extraction de chaînes relatives aux opinions et apprentissage

Le système extrait les mots et chaînes des critiques avec la note de cette critique pour créer une association (mots/chaînes - poids). Le poids étant calculé à partir de la fréquence d'apparition du terme dans sa catégorie. Ce poids sera la base de notre catégorisation.

Une fois les modèles de termes extraits, un poids est calculé dans chaque classe et pour chaque critique. La critique se verra attribuer la catégorie pour laquelle elle obtient le poids le plus fort. Nous avons effectué la catégorisation de textes avec le classifieur CosinusGini. Il s'agit d'un classifieur fondé sur la mesure de similarité cosinus entre un vecteur W_d du document d à classer et un vecteur W_c représentant la catégorie c dont les dimensions sont les TF - IDF [2] des mots ou chaînes i les composant. Les valeurs attribuées à chaque mot (ou chaîne) incluent en plus une variable gini (2), nommée critère de pureté de Gini introduite dans un tel contexte applicatif par Torres-Moreno et al. (2012). [3]

Le cosinus se calcule comme présenté sur la formule (1) :

$$cos(d,c) = \frac{\sum\limits_{i \in d \cap c} \omega_{i,d} \times \omega_{i,c}}{\sqrt{\sum\limits_{i \in d} \omega_{i,d}^2 \times \sum\limits_{i \in c} \omega_{i,c}^2}} \tag{1}$$

où $\omega_{i,d}$ le poids du mot *i* dans le document *d* = $TF_{i,d} \times log(\frac{N}{DF_i}) \times gini_i$
et où :

$$gini = 1 - \sum_{c=1} P\left(\frac{c}{i}\right)^2 \tag{2}$$

Le système devra également décider (lorsque cela est nécessaire) d'utiliser une chaîne plutôt qu'un seul terme, car les performances peuvent se retrouver grandement améliorées dans certains cas Lavalley et al. (2009, 2010). Des exemples de chaîne à privilégier sont « mise-en-scène » , « meilleur-film » ou encore « excellent-film » ce qui permet de compter la fréquence d'apparition de l'agglutination plutôt que celle de chaque terme. La chaîne aura ainsi un pouvoir discriminant supérieur, augmentant son influence sur la décision finale. Ces agglutinations évitent également que les négations ne dégradent le pouvoir discriminant d'un terme par exemple les « pas-bon » , « pas-fameux » ou dans le sens inverse le « pas mauvais » . Une méthode qui a été utilisée pour calculer les agglutinations se base sur le coefficient de Gini. Deux termes sont agglutinés lorsque la valeur de Gini de l'agglutination est plus importante que celle des deux termes séparés. Cela permet d'agglutiner uniquement les éléments discriminants. Si gini(mot_1) et gini(mot_2) < gini(mot_1-mot_2).

2. (Term Frequency - Inverse Document Frequency Salton et Buckley (1988). Dans nos travaux, le TF représente le nombre de documents contenant (au moins une fois) le terme i dans la catégorie c normalisé par le nombre de documents dans la catégorie. IDF représente le logarithme de l'inverse de la proportion de (μC) du corpus qui contiennent (au moins une fois) le terme i.

3. Sa valeur représente la dispersion du terme dans l'ensemble des catégories il s'agit de la somme au carré du nombre de (μC) contenant le terme dans chaque catégorie divisé par le nombre total de documents contenant le terme. Son rôle consiste à réduire l'influence des mots ayant un faible pouvoir discriminant (ceux qui apparaissent de manière équi répartie dans les différentes classes). Sa valeur est ainsi maximale quand le mot n'apparaît que dans une seule catégorie et minimale quand il apparaît équitablement dans toutes les catégories.

Bien que les classifieurs de type mesure de similarité cosinus ne soient pas réputés pour être les plus performants, ils ont une faible complexité et ne nécessitent pas d'ajustement de paramètres. Un autre élément important est que ce type de système nous permet d'accéder aux critères discriminants et d'analyser les erreurs de classification.

3 Protocole expérimental

3.1 Catégorisation de textes en utilisant les chaînes extraites

Le corpus dont nous disposons pour mener à bien nos expériences provient du portail communautaire sur le cinéma Vodkaster [4]. Ouverte au dernier trimestre 2009, la plateforme Vodkaster a été développée sur un concept totalement novateur en Europe à cette époque. Vu comme une plateforme de diffusion d'extraits de film, le site permet de découvrir des films via des extraits (généralement les scènes les plus marquantes ou célèbres d'un film). Cependant Vodkaster s'est démarqué en permettant aux utilisateurs de participer à l'enrichissement de la base d'extraits, mais aussi de critiquer les œuvres.

Le problème de la catégorisation se matérialise la prédiction d'une note qu'un utilisateur du site Vodkaster aurait pu attribuer à un film, prédiction faite en fonction du contenu de la micro-critique associée au couple (utilisateur, film). Contrairement aux autres systèmes de prédiction existants, ici la note n'est plus l'élément central de la critique. Néanmoins, sa présence permet, lors de la phase d'apprentissage, de pouvoir affecter un score (positif ou négatif) à un terme (ou une chaîne) porteur d'opinion. La note devient le point d'appui de l'extraction du contenu de la critique à laquelle elle est associée.

3.2 Catégorisation de textes : prise en compte des métadonnées

Nous disposons avec le corpus de données (voir Section 4) de nombreuses informations supplémentaires concernant la critique :

- l'identité de l'auteur de la critique (son pseudo sur le portail) ;
- le titre du film dont il est question dans la critique ;
- la date d'émission de la critique.

Afin de tenir compte du comportement passif de l'utilisateur mais également de celui du film, nous avons décidé d'inclure les deux premières données dans le système. Cela se matérialise par la prise en compte dans la critique du pseudo et du titre du film comme chaîne dans la critique au même titre que les termes qui la compose. De lui même, le système va intégrer dans le modèle les éléments rajoutés.

Nous testons également des combinaisons temporelles en utilisant dans notre corpus de test des critiques plus récentes que celles utilisées dans le corpus d'apprentissage. Même si globalement prédire la note d'une critique ancienne à partir de critiques récentes ne parait pas pertinent, la prédiction pour des critiques récentes est différente selon si l'on prend en compte des critiques proches ou lointaines.

4. http://www.vodkaster.com/

4 Corpus

Le corpus dont nous disposons est un ensemble de micro-critiques (μC), d'utilisateurs du site Vodkaster. Une μC est un tuple (utilisateur, date, film, note, critique).

Chacune des μC est structurée en six champs de la façon suivante :

```
 1323556 peonidelavega 05/01/12-13h29 Le Pacte 0.5 Chef d'oeuvre!
non je plaisante...
```

- L'échelle des notes comporte dix barreaux de 0,5 à 5.
- La critique est dite micro-critique car d'une longueur maximale de 140 caractères comme pour les tweets.

L'identifiant est unique et correspond au numéro de la critique dans la base de données du site (et dans le corpus). La répartition des critiques en fonction des notes dans l'ensemble des sous corpus est représentée en figure 1 :

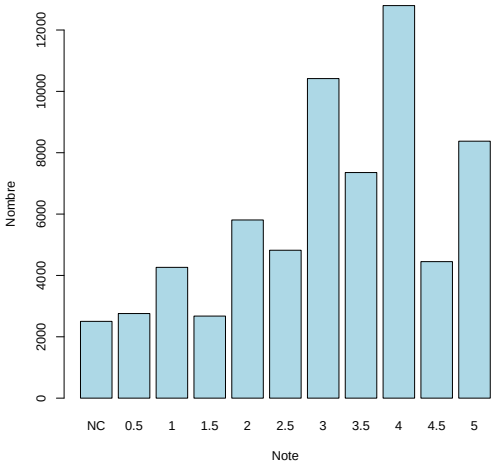

FIG. 1 – *Répartition des critiques en fonction des notes*

Il est utile de dresser quelques constats :

- Sur un total de 66 200 micro-critiques, 2 500 sont dépourvues de note ;
- La moyenne des notes sur 66 200 μC est de 3,2 (la moyenne de l'échelle étant 2,5) ;
- Les notes intermédiaires (X,5) sont moins utilisées ;
- Les internautes critiquent majoritairement de manière positive les œuvres ;

Les répartitions des critiques par utilisateur ou par film sont assez inégales. Ceci est un élément préjudiciable au processus de recommandation dans le sens ou la majorité des films n'est que peu critiquée et une majorité d'utilisateurs ne critique que peu de films.

Pour nos expériences, nous avons divisé le corpus en trois parties respectivement apprentissage (taille variable entre 15 000 et 34 000), développement (6 000 μC) et test (6 000 μC). Ce découpage a été fait tout en respectant l'ordre chronologique des critiques afin d'éviter un entre-laçage [5].

La note affectée par les internautes ne se base pas pour tous sur les mêmes critères. De ce fait les nuances entre le « 3/5 » d'une personne et le « 3,5/5 » d'un autre internaute sont tellement infimes que le système serait très certainement incapable de déterminer une note correcte dans ce cas-là.

De ce fait l'échelle de note a été ramenée à deux barreaux [6] lors des expériences :

– Positif ;
– Négatif.

Se baser sur une recherche de tranche et non d'une note permet de compenser le problème de l'unicité de l'échelle de valeur. Pour limiter d'autant plus ce problème, les critiques neutres seront exclues de nos expériences.

Le tableau 1 illustre un exemple de ce problème pour un des films les plus critiqués du corpus : X-Men. Des termes *a priori* élogieux, sont utilisés avec une grande variabilité de note :

terme	moyenne	min	max	fréquence
meilleur	4,05	3	5	9
bon	3,55	2	5	9
bien	3,11	1	4	9
très-bon	3	2	4	3

TAB. 1 – *Dispersion des termes « positifs » pour le film X-Men.*

Une opération de nettoyage du corpus d'apprentissage est réalisée : le système évalue un corpus de test identique au corpus d'apprentissage. Il s'agit de prédire une note sur une critique que l'on a déjà rencontrée durant l'apprentissage. Les erreurs d'évaluation sont considérées comme du bruit et donc éliminées du corpus d'apprentissage pour les expériences suivantes. Cela permet d'éliminer le bruit du corpus d'apprentissage. Ces expériences ont été menées sur un corpus ayant subit les modifications suivantes :

– Des agglutinations spécifiques aux classes ont été calculées ;
– Les couples doublons (utilisateur,film) ont été éliminés, seule la μC la plus récente a été conservée en cas de double ;
– Un anti-dictionnaire a été utilisé ;

Le corpus après traitement contient 41 126 μC. Voici quelques caractéristiques importantes :

– Il y a 1 723 contributeurs différents soit en moyenne 24 critiques par contributeur ;
– 510 utilisateurs n'ont émis qu'une seule critique ;
– Le plus gros contributeur a posté 1559 critiques depuis son inscription à l'ouverture de la plateforme ;
– 7 415 films ont été critiqués soit en moyenne 5 critiques par film ;
– 3 400 films n'ont été critiqués qu'une seule fois ;

5. L'entre-laçage est un phénomène qui apparaît lorsque l'on rencontre dans les tests des éléments déjà appréciés au moment de l'apprentissage. Ce phénomène biaise les résultats pour des données où les utilisateurs se répondent les uns aux autres sur un sujet donné. Tout sujet croisé dans l'apprentissage qui est recroisé dans le test est de fait reconnu et cela facilite de façon artificielle la tâche de classification. La notion temporelle fera l'objet d'explication supplémentaire par la suite

6. Sont considérées neutres les critiques dont la note est comprise entre 2 et 4 (non inclus) ces critiques dont la note est proche de la moyenne. Un traitement différent entre un message réellement neutre et un message nuancé pourrait être envisagé.

– Le film (Drive) le plus critiqué l'a été 241 fois depuis sa sortie ;

5 Résultats

Les résultats sont exprimés en pourcentage de réussite. Ce pourcentage correspond au Fscore définit sur la formule (3).

$$Fscore = \frac{Precision \times Rappel}{Precision + Rappel} \tag{3}$$

La précision pour une classe Pc est définie par :

$$Pc = \frac{\text{Nombre de documents correctement attribués à la classe } c}{\text{Nombre de documents attribués à la classe } c} \tag{4}$$

Rappel Rc pour une classe

$$Rc = \frac{\text{Nombre de documents correctement attribués à la classe } c}{\text{Nombre de documents appartenant à la classe } c} \tag{5}$$

Les résultats présentés au tableau 2 proviennent d'expériences menées dans les conditions suivantes :
– Apprentissage composé de 33 922 μC (21 198 positives et 12 724 négatives)
– Test composé de 6 000 μC (3 750 positives et 2 250 négatives)
– Un écart de 15 jours entre l'ensemble d'apprentissage et de test (soit 1 204 critiques)

Les résultats du système de classification avec deux stratégies d'apprentissage sont présentés au tableau (2).

Méthode	Fscore
avant l'ajout des métadonnées	0.8298
après l'ajout du pseudo	0.8405
après l'ajout du titre	0.8523
après l'ajout des deux	**0.8585**

TAB. 2 – *F-scores obtenus sur la catégorisation du corpus de Test*

L'ajout de métadonnées dans la critique semble avoir une influence très positive sur l'opération de classification. Toutefois, comme nous avons pu le constater, toutes les métadonnées ne se valent pas car l'ajout du titre du film offre une amélioration plus significative que l'ajout du pseudo.

Ce qui amène au constat suivant, le système de classification semble avoir un problème bien connu dans le domaine de la recommandation de contenu : le démarrage à froid. Ici, cela se traduit par de mauvais résultats sur de nouveaux films.

L'analyse de nos 849 erreurs nous permet d'obtenir plus d'informations :

En positif	Nombre d'erreurs	Taux d'erreurs
posteur inconnu film connu	105	12,36
posteur connu film inconnu	112	13,19
film inconnu posteur inconnu	40	4,71
film connu posteur connu	143	16,84
En négatif		
posteur inconnu film connu	76	8,95
posteur connu film inconnu	179	21,08
film inconnu posteur inconnu	34	4
film connu posteur connu	160	18,8

TAB. 3 – *Taux d'erreurs*

Le tableau (3) montre que le taux d'erreurs est plus important pour les critiques dont le film n'a jamais été croisé dans l'apprentissage. Et dans ce cas, l'apport de la connaissance de l'utilisateur émettant la critique n'aide pas le système. Comme évoqué précédemment, prédire la note d'une critique ancienne à partir de critiques récentes ne parait pas pertinent. Toutefois, la prédiction pour des critiques récentes avec des critiques anciennes présente un paradoxe. En effet au delà d'un certain seuil, augmenter la masse d'apprentissage par retour dans le passé n'améliore pas les performances. L'influence observée est parfois même négative. A taille égale les données d'apprentissage les plus proches dans le temps des données de test offrent même de meilleurs résultat pour une période de test donnée comme le montre la figure 2 :

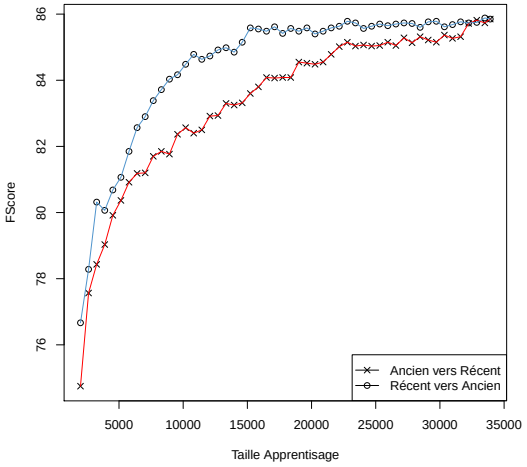

FIG. 2 – *Évolution du Fscore en fonction de la taille d'apprentissage*

D'autres tests ont été menés afin de détecter un taille ou segment d'apprentissage optimal. Les tests ont montré que pour chaque sous corpus de test il existe un sous corpus d'apprentissage qui donne de meilleurs résultats que le corpus d'apprentissage considéré dans son ensemble.

Il convient toutefois de relativiser, les critiques de cinéma sont liées à (voire influencées par) des phénomènes d'actualité (pics d'activités durant les festivals, morosité après une élection ou faits divers tragiques). Les données ne disposent pas d'une profondeur temporelle suffisante pour asseoir de manière fiable ces résultats.

6 Perspectives et discussions

La méthode décrite précédemment permet d'extraire des chaînes de mots relatives aux différentes opinions exprimées. Ces chaînes offrent la possibilité de visualiser des expressions exprimant une opinion.

Concernant les critiques neutres, des progrès restent encore à faire afin de pouvoir distinguer des critiques neutres (sans avis tranché) de critique nuancées (relevant des aspects positifs contre balancés par des aspects négatifs). Un effort doit être fait sur la détection de la polarité de certains idiomes, afin de mieux prendre en compte les renversements effectués par certains marqueurs de négations mais également par les marqueurs d'ironie et langage familier parfois employés en fin de critique pour tenter d'éviter les erreurs vues précédemment. Cela pourrait s'envisager en travaillant par exemple avec des marqueurs grammaticaux ou un système de patrons. Voici une μC illustrant parfaitement la chose :

1323556 peonidelavega 05/01/12-13h29 Le Pacte 0.5 Chef d'oeuvre ! non je plaisante...

Cela amène également à la détection d'utilisateurs systématiques faisant preuve d'un grand laxisme ou d'une grande sévérité envers un acteur, un réalisateur ou un genre, afin de mieux pondérer le contenu de la critique mais aussi d'affiner les recommandations faites. Ainsi il sera possible de proposer un système de résumés où seront mis en avant les points qui ont été réellement appréciés ou non dans le film et ceux qui seraient susceptibles d'intéresser l'utilisateur à qui s'adresse la recommandation. De même il faudra également être capable de mieux différencier les critiques de gros ou petits utilisateurs (en nombre de critiques). Ceci permettra d'affecter un poids pouvant améliorer la crédibilité d'un avis (par exemple dans le cas d'un compte créé uniquement pour influencer les avis sur un film).

Références

Lavalley, R., P. Bellot, et M. E.-B. M (2009). Interactions entre le calcul de collocations et la catégorisation automatique de textes. *Actes de CORIA*, 251–265.

Lavalley, R., C. Clavel, et P. Bellot (2010). Extraction probabiliste de chaînes de mots relatives à une opinion. *TAL 51*, 101–130.

Meyer, F., E. Gaussier, F. Clerot, et J. Schluth (2011). Apport des données thématiques dans les systèmes de recommandation hybridation et démarrage à froid. *EGC*.

Salton, G. et C. Buckley (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management 24*, 513–523.

Torres-Moreno, J.-M., M. El-Beze, P. Bellot, et F. Bechet (2012). *Opinion detection as a topic classification problem, Chapitre 9 Textual Information Access*. Avignon FRANCE : ISTE Ltd John Wiley and Son.

L'importance des métadonnées

Summary

The presented research focuses on cinema through data and metadata provided by the site vodkaster (movies social network). In order to feed a reviews-based Recommendation System, our first goal was to design a machine learning system to predict the category (positive or negative) of reviews witten by a Vodkaster member. This prediction is based on the words contained in the review only. The reviews are called micro-reviews (μC) because it shares the same format than a tweet (140 characters). In the prediction process, we take into account the identity of the user (his nickname) and the film's title, both integrated as words in the μC's word bag. This additional step has improved the overall performance of the system whatever the method used for prediction. We then focus on the optimal learning-segment detection. We show that with an equal number of documents, the closest learning segment (temporally speaking) to test data gives better results than documents more distants in time.

Index