HAL
open science

Actes de l'atelier Journalisme Computationnel 2017

Laurent Amsaleg, Vincent Claveau, Xavier Tannier

To cite this version:

Laurent Amsaleg, Vincent Claveau, Xavier Tannier (Dir.). Actes de l'atelier Journalisme Computationnel 2017. 2017. hal-01643634

HAL Id: hal-01643634

https://hal.science/hal-01643634v1

Submitted on 21 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Atelier Journalisme Computationnel

Organisateurs : : Laurent Amsaleg (IRISA-CNRS), Vincent Claveau (IRISA-CNRS), Xavier Tannier (LIMSI-Univ. Paris Sud)

PRÉFACE

Motivations et objectifs

Loin de l'image des autoroutes de l'information, l'espace numérique tient plutôt des chemins tortueux dans lesquels les professionnels de l'information doivent rechercher, filtrer, croiser, vérifier ou décoder. Les volumes de données manipulés, leur variété (vidéos, textes, images, bases de connaissances...) et leur vélocité offrent des opportunités pour appréhender l'information autrement, mais posent aussi de nombreux problèmes de recherche désormais rangés sous l'étiquette *Big Data*. Dans ce contexte, journalistes et technologues ont développé la notion de journalisme de données. Cette pratique nouvelle du journalisme tire partie des données numériques disponibles pour produire et distribuer l'information. Elle bénéficie notamment de la popularité croissante de l' *Open Data*, du développement de bases de connaissances structurées, du traitement automatique des langues, ainsi que des travaux récents en visualisation de données, pour faciliter l'analyse de l'information et proposer une grande variété de points de vue.

Certains journalistes utilisent des outils visant à améliorer leur productivité ou leur couverture d'un sujet (bases de connaissances, réseaux sociaux...). D'autre part, les chercheurs en TAL, RI, BD, IA utilisent massivement le matériel journalistique dans leurs travaux : articles de presse, dépêches d'agence, images, vidéos. Récemment, plusieurs projets de recherche sur ces thèmes et impliquant des organes de presse ont vu le jour. Une journée organisée à l'IRISA Rennes en mars 2016 a montré l'intérêt de nombreux organismes de recherche, entreprises privées et professionnels des médias sur ce sujet[1].

L'objectif principal de l'atelier est de servir de lieu de rencontre entre les différents acteurs de cette communauté naissante. Ceux-ci relèvent souvent de sous-domaines différents de l'informatique, se rencontrant assez peu, alors que les problématiques impliquent une démarche intégrant tous ces sous-domaines. La constitution d'un panorama des travaux, l'éventuel partage d'outils, données, *benchmarks* ou de résultats pourront enrichir cette réflexion.

Un autre objectif de cet atelier est de mieux tenir compte de la réalité du travail journalistique. Cela part du constat qu'entre chercheurs et professionnels de l'information, il reste difficile de pérenniser les collaborations et de développer des outils permettant de travailler plus efficacement avec les masses de données, outils qui seraient utilisés en aide à la production éditoriale quotidienne. Cet atelier a pour but de stimuler la réflexion et la discussion sur les bénéfices concrets que les journalistes peuvent retirer des outils développés par les spécialistes des STIC, sur les effets que ceux-ci peuvent avoir sur la pratique journalistique, et sur les nouvelles analyses liées à l'exploitation des médias. Si les informaticiens, au sens large, proposent des outils aux professionnels de l'information, ces derniers ont aussi à exprimer leurs besoins, leurs attentes, à partager leurs manières de procéder. Les disciplines informatiques concernées peuvent alors se pencher sur ces usages inédits, demandant de résoudre des problèmes de recherche durs, exigeant de se poser des questions tant d'ordre

[1] http://compjournalism2016.irisa.fr

méthodologique que plus appliqués où il pourrait être question d'adapter des techniques partiellement existantes à ces nouveaux contextes ou d'en inventer de nouvelles. L'atelier a donc pour but de faire circuler les idées tant des journalistes vers les informaticiens que des informaticiens vers les journalistes.

Autour de cette volonté de croiser les discours, de faire des allers-retours entre journalistes et informaticiens, nous avons bâti un appel a communications dont le cœur est constitué des thèmes suivants :

- la détection d'événements,
- le *fact-checking*, le décodage,
- les études sociologiques ou historiques,
- la fiabilité des sources,
- l'exploration d'archives d'actualités,
- la génération automatique de contenu journalistique,
- la visualisation de données, la navigation dans de grandes masses de données,
- la production participative *(crowdsourcing)* pour le journalisme,
- la dissémination des nouvelles à travers les réseaux sociaux,
- les outils "intelligents" pour les journalistes,
- la recommandation, la personnalisation,
- la détection de plagiat, de cliché, de biais, de propagande, de fausses informations *(hoax)* dans le texte, les images ou les vidéos,
- l'analyse du discours politique,
- la contextualisation de l'information,
- la diversité des sources.

Programme

Les articles reçus en réponse à l'appel ont été relus par un comité de lecture (liste ci-après) incluant informaticiens et professionnels de l'information. Ils nous ainsi permis de bâtir un programme varié autour de trois exposés longs et cinq exposés courts abordant des points techniques, livrant des analyses sociologiques, ou témoignant des usages journalistiques d'outils informatiques.

Exposés longs :

- Julien Velcin, Jean-Claude Soulages, Solange Kurpiel, Luis Otavio, Myrian Del Vecchio and Frédéric Aubrun. Fouille de textes pour une analyse comparée de l'information diffusée par les médias en ligne : une étude sur trois éditions du Huffington Post
- Jeremy Vizzini, Cyril Labbé and François Portet. Génération automatique de billets journalistiques : singularité et normalité d'une sélection
- Marie-Luce Viaud, Nicolas Hervé and Julia Cagé. Analyse des Media Français: Quand l'économie rencontre la fouille de donnée

Exposés courts :

- Béatrice Mazoyer, Nicolas Turenne and Marie-Luce Viaud. Étude des influences réciproques entre médias sociaux et médias traditionnels
- Nicolas Médoc, Mohammad Ghoniem and Mohamed Nadif. Analyse exploratoire de corpus textuels pour le journalisme d'investigation
- Natalia Grabar and Mason Richey. Détection automatique de grandes thématiques de la propagande Nord Coréenne
- Julien Maitre, Michel Menard, Guillaume Chiron and Alain Bouju. Utilisation conjointe LDA et Word2Vec dans un contexte d'investigation numérique
- Guillaume Chiron, Jean-Philippe Moreux, Antoine Doucet, Mickael Coustaty and Muriel Visani. Erreurs OCR et biais d'indexation : impact sur les usages

Pour compléter cette sélection, nous avons invité Gauthier Bravais, Pierre Bellon et Lucas Piessat, de l'agence Skoli, pour présenter leurs travaux sur l'analyse du traitement médiatique de faits d'actualités mêlant utilisation d'outils de fouille et rendu grand-public. Leur présentation est intitulée : Une analyse de données textuelles des archives numériques de la presse française pour explorer le traitement médiatique de l'Islam. L'exemple d'une collaboration chercheur / agence Web spécialisée.
Dans cette présentation, G. Bravais et ses collègues montrent comment l'agence Skoli s'est associée avec Moussa Bourekba (chercheur CIDOB, Barcelone) pour étudier le traitement médiatique de l'Islam en France (1997-2015). Leur collaboration, originale à ce niveau, s'est articulée autour d'une analyse de données textuelles de milliers d'articles issus des archives numériques de trois quotidiens français de référence (Le Monde, Le Figaro, Libération) et de la réalisation d'une interface Web de restitution mêlant datavisualisations et décryptages.

Laurent Amsaleg — IRISA - CNRS
Vincent Claveau — IRISA - CNRS
Xavier Tannier — LIMSI - Univ. Paris-Sud

Membres du comité de lecture

Le comité de lecture est constitué de :

Laurent Amsaleg, IRISA-CNRS
Max Chevalier, IRIT - Univ. Toulouse
Vincent Claveau, IRISA-CNRS
Géraldine Damnati, Orange Labs
Claude de Loupy, Syllabs
Michel Le Nouy, Ouest-France
Damien Nouvel, ERTIM-INALCO
Xavier Tannier, LIMSI - Univ. Paris Sud
Raphael Troncy, EURECOM
Julien Velcin, ERIC - Univ. Lyon 2
Haïfa Zargayouna, LIPN - Univ. Paris 13

TABLE DES MATIÈRES

Exposés longs

Fouille de textes pour une analyse comparée de l'information diffusée par les médias en ligne : une étude sur trois éditions du Huffington Post
Julien Velcin, Jean-Claude Soulages, Solange Kurpiel, Luis Otavio, Myrian Del Vecchio, Frédéric Aubrun 1

Génération automatique de billets journalistiques : singularité et normalité d'une sélection
Jérémy Vizzini, Cyril Labbé, François Portet 13

Analyse des Media Français : Quand l'économie rencontre la fouille de donnée
Marie-Luce Viaud, Nicolas Hervé, Julia Cagé 25

Exposés courts

Étude des influences réciproques entre médias sociaux et médias traditionnels
Béatrice Mazoyer, Nicolas Turenne, Marie-Luce Viaud 37

Analyse exploratoire de corpus textuels pour le journalisme d'investigation
Nicolas Médoc, Mohammad Ghoniem, Mohamed Nadif 41

Détection automatique de grandes thématiques de la propagande Nord Coréenne
Natalia Grabar, Mason Richey 45

Utilisation conjointe LDA et Word2Vec dans un contexte d'investigation numérique
Julien Maitre, Michel Menard, Guillaume Chiron, Alain Bouju 57

Erreurs OCR et biais d'indexation : impact sur les usages
Guillaume Chiron, Jean-Philippe Moreux, Antoine Doucet, Mickael Coustaty, Muriel Visani 69

Index des auteurs **75**

Fouille de textes pour une analyse comparée de l'information diffusée par les médias en ligne : une étude sur trois éditions du Huffington Post

Julien Velcin*, Jean-Claude Soulages**, Solange Kurpiel**,****,
Luis Otávio Dias****, Myrian Del Vecchio**,****, Frédéric Aubrun***

* Université de Lyon (Université Lyon 2, ERIC EA 3083)
Julien.Velcin@univ-lyon2.fr
http://mediamining.univ-lyon2.fr/velcin
** Université de Lyon (Université Lyon 2, Centre M. Weber)
Jean-Claude.Soulages@univ-lyon2.fr
solange.kurpiel@gmail.com
*** Université de Lyon (Université Lyon 2)
aubrunf@gmail.com
**** Université Fédérale du Paraná (Brésil)
myriandel@gmail.com
fototavio@yahoo.com.br

Résumé. Cet article présente un processus d'analyse mis en place dans le cadre d'une collaboration entre des chercheurs en informatique, en sociologie et en sciences de l'information et de la communication, à l'occasion du projet Journalisme A l'heure Du Numérique. Le processus consiste pour le moment en un recodage manuel de thématiques extraites de manière totalement non supervisée à partir des données textuelles publiées sur le site du Huffington Post. Cette démarche rend possible une analyse comparée d'un corpus d'articles publiés durant l'été 2016 dans trois éditions différentes du journal (française, américaine, brésilienne). Les premiers résultats présentés permettent de valider la démarche tout en interrogeant sur les améliorations possibles, en particulier une automatisation plus importante des étapes qui composent le processus.

1 Introduction

L'arrivée des nouvelles formes de communication virtuelle, qu'il s'agisse des blogs ou des médias sociaux, a révolutionné la manière dont l'information est publiée et diffusée. De nombreux métiers s'en trouvent bouleversés, comme celui de journaliste (Charon et Papet, 2014). La diversité de ces nouveaux lieux d'expression et leur caractère hautement concurrentiel a multiplié les points de vue sur l'actualité, ce qui conduit à nous interroger sur les choix qui y sont faits, que ceux-ci soient explicites (politique éditoriale) ou implicites (culture locale). Cette interrogation se pose avec d'autant plus d'acuité lorsque l'on observe l'information diffusée par un même média global, comme c'est le cas du Huffington Post abordé dans ce tra-

vail, développant des stratégies éditoriales locales (version américaine, française, brésilienne, arabe, etc.). Au-delà de la tension entre les stratégies éditoriales globales et locales, déjà mise en lumière par le biais de la publicité (Aubrun, 2015), d'autres critères peuvent intervenir à différents degrés dans le choix final qui sera réalisé, comme par exemple le coût élevé de la création d'un reportage original par rapport à la reproduction *ad nauseam* d'articles publiés ailleurs. Cet article présente un travail pluridisciplinaire pour capturer ces différences à l'aide d'outils de fouille de données et en faire une première analyse critique.

Afin d'être en mesure d'expliciter les biais dans l'information véhiculée par les médias, il est nécessaire de parvenir à comparer l'information à partir d'un référentiel commun malgré une différence parfois très importante dans les sujets traités et la langue employée. Dans ce travail, nous proposons d'articuler une méthode automatique qui extrait des thématiques de manière non supervisée directement à partir des données, spécifiques à chacun des corpus étudiés, avec un recodage à l'aide de catégories transversales, pour le moment réalisé de manière manuelle par des sociologues. L'intérêt est double : a) conserver une certaine flexibilité en permettant de ne pas être trop collé aux catégories proposées par les sites des médias, et donc permettre à des catégories nouvelles d'apparaître, b) garantir un ensemble de catégories de qualité validées par les experts et permettant de rendre les corpus comparables. Il s'agit donc d'une manière de construire un résumé associé à chaque média, et plus précisément dans notre cas à chaque version du même média, et ce de manière semi-automatique.

Ce travail représente la première étape du projet collaboratif Journalisme A l'heure Du Numérique [1] (JADN) qui mobilise des sociologues, spécialistes de l'analyse des médias, et des informaticiens, spécialistes de fouille de données. Une première application visée est de fournir aux journalistes ou aux chercheurs un outil permettant de visualiser la couverture de l'actualité. Un deuxième enjeu est d'ordre culturel ou géo-culturel puisqu'il s'agit de mettre au jour les points de convergence ou de divergence dans les différentes éditions d'un même média (ici, le Huffington Post). Les tout premiers résultats obtenus sur un corpus de 18 555 articles de presse, publiés sur le site du Huffington Post en trois langues (anglais, français, portugais) et sur une durée d'environ trois mois durant l'été 2016, sont présentés dans cet article. Ils permettent de donner un aperçu du "paysage informationnel" (voir Appadurai (2011) pour la notion de *mediascape*) des trois corpus et d'en réaliser une première analyse comparée. Outre une brève interprétation qualitative de ces résultats, ils mettent en lumière différentes problématiques que nous souhaitons développer dans nos travaux futurs. Nous prévoyons en particulier d'automatiser une grande partie de ce processus, encore très manuel, telles que la construction des catégories et la comparaison des résumés.

La suite de cet article est divisée en quatre parties. La première section décrit l'approche systémique que nous avons suivie, dans laquelle nous combinons des outils d'apprentissage automatique avec une analyse qualitative. La deuxième section présente les données récoltées, donnant une brève motivation concernant l'étude d'un média en particulier, avant de décrire le protocole expérimental et les premiers résultats obtenus. La troisième section donne quelques éléments bibliographiques, à la fois en analyse comparée des médias et en fouille de données à base de modèles thématiques. Enfin, la dernière section est consacrée à la conclusion et aux perspectives de ce travail.

1. http://jadn.univ-lyon2.fr

2 Approche proposée

La démarche décrite ci-dessous est le fruit d'un consensus qu'il a fallu trouver entre les deux équipes, comme c'est souvent le cas dans les projets pluridisciplinaires. Elle repose principalement sur l'idée de pouvoir remettre en question chacune des étapes, par exemple le nombre de thématiques choisies ou les catégories construites pour subsumer ces thématiques. Elle est composée de cinq étapes décrites ci-dessous :

— Constitution de plusieurs corpus composés d'articles publiés durant la même période. A ce stade, aucune hypothèse n'est formulée et l'intégralité des articles est récupérée pour les analyses ultérieures.

— Extraction automatique des thématiques en adoptant une approche de bas vers le haut, c'est-à-dire totalement non supervisée. A ce stade, plusieurs modèles sont proposés dans la littérature (voir section 4). Dans notre cas, nous avons choisi un modèle reconnu et pour lequel il convient de fixer le nombre de thématiques attendu (modèle LDA de Blei et al. (2003)). L'avantage est que l'on peut s'attendre à obtenir des thématiques de granularité comparable, même si ce ne sera pas toujours le cas comme nous l'avons constaté lors des expérimentations.

— Détection des thématiques jugées erronées car trop difficiles à interpréter. Cette détection se fait uniquement sur la base des 10 mots-clefs les plus importants. Il peut s'agir, par exemple, de thématiques regroupant des mots outils spécifiques au corpus ou des erreurs dans l'acquisition des données (par exemple, nous avions une thématique regroupant des balises javascript non détectées par l'algorithme d'acquisition).

— Attribution d'une catégorie à chaque thématique conservée, plusieurs thématiques pouvant partager la même catégorie (par exemple "politique étrangère" pour une thématique sur le coup d'état en Turquie ou sur le Brexit). Les catégories sont fixées par le sociologue à la lumière de la lecture des thématiques, pour le moment fournies sous la forme d'une liste des 10 mots-clefs les plus importants (donc ceux qui maximisent la probabilité $p(w/z)$, w étant un mot du dictionnaire et z la thématique en question). Il a été question de fournir les articles les plus centraux, donc ceux qui maximisent $p(d/z)$, mais il a été décidé de ne pas utiliser ce surcroît d'information pour le moment et d'étudier si les mots en tête de liste pouvaient s'avérer suffisants dans la compréhension du contenu abordé par la thématique. Pour fixer les catégories, la méthode suivante a été appliquée :

1. Sur la base des 10 mots-clefs d'une thématique, le spécialiste choisit une catégorie qui lui semble assez générale pour couvrir plusieurs thématiques (le flou sur ce que recouvre le terme "générale" est conservé à dessein). A ce stade, celui-ci n'est pas restreint et peut parfaitement créer de nouvelles catégories.
2. Le spécialiste peut revenir sur l'attribution de catégories à des thématiques précédentes, voire supprimer certaines catégories si celles-ci s'avèrent inappropriées (la catégorie "élections américaines" a finalement été exclue du corpus US).
3. Le spécialiste modifie les catégories des différents corpus afin de converger vers une unique liste de catégories qui permet d'encoder chaque thématique.
4. Il est possible à présent de compter, pour chaque corpus, le nombre de thématiques et d'estimer le volume d'articles pour chacune des catégories. Pour une catégorie donnée c, on connaît le nombre de thématiques associées et on peut estimer son

importance en calculant $f(c) = \sum_{z \in Z_c} \sum_{d \in D} p(z/d)$, où Z_c est l'ensemble des thématiques pour la catégorie visée c et d est un document de notre corpus D. On peut interpréter cette valeur comme le nombre d'articles publiés dans cette catégorie, bien qu'il s'agisse d'une estimation étant donné qu'un document est associé à plusieurs thématiques dans ce genre de modèle. Dans les résultats présentés dans cet article, nous avons arrondi cette estimation à l'entier le plus proche afin d'être cohérent avec cette interprétation que nous en donnons.

— Puisque les différents corpus sont codés à l'aide du même ensemble de catégories, il est à présent possible de les comparer sur la base des différentes valeurs de $f(c)$, comme nous le verrons dans le diagramme de la figure 3.

3 Expérimentations

3.1 Données

Dans ce travail, nous avons choisi de nous concentrer sur des articles de presse publiés sur le site du Huffington Post[2]. Ce média a l'avantage de répondre à un ensemble de préoccupations méthodologiques et thématiques en lien avec l'évolution du journalisme à l'ère du numérique. The Huffington Post est un *pure player* gratuit fondé aux États-Unis par Ariana Huffington, Kenneth Lerer et Jonah Peretti en 2005, racheté par AOL pour 315 millions de dollars américains en 2011 et décliné en 12 éditions à travers le monde, dont les éditions française (23 janvier 2012), et brésilienne (29 janvier 2014). Présenté à la presse et au public comme un média indépendant qui vise à révolutionner le journalisme grâce à une offre numérique « alternative » du point de vue technologique comme politique, le Huffington Post a fait de sa structure algorithmique, de son modèle économique comme de son panel de rédacteurs triés sur le volet, les leviers de son succès américain puis international. A mi-chemin entre un agrégateur d'informations et un producteur de nouvelles, il compte dans ses rangs des journalistes salariés rompus au travail de curation et des contributeurs extérieurs bénévoles, jouissant d'une grande notoriété. L'originalité de son modèle éditorial repose sur un spectre très large d'articles d'actualité mais surtout sur des articles d'opinion sous forme de blogs signés par des experts ou des personnalités connues comme Barak Obama, David Cameron, Michael Moore, Rachida Dati, Madona, etc. Son positionnement à la fois global et local constitue une formidable opportunité pour étudier le poids de la globalisation dans la production de l'information à travers la distribution et la reprise de nouvelles et de blogs par chacune des éditions nationales.

Afin d'étudier ces données, nous avons mis en place un aspirateur de flux RSS sur quatre versions du Huffington Post (américaine, française, brésilienne, arabe) à partir de juin 2016, à la fois sur les articles de presse que sur les articles de blog. Chaque article est associé à un titre, au corps de l'article, à une date et à un auteur. Les résultats présentés dans cet article correspondent à presque trois mois de collecte (plus précisément du 20 juin au 8 septembre 2016) et se concentrent uniquement sur les trois premières versions et sur les articles de presse. Le tableau de la figure 1 donne des statistiques simples sur les corpus étudiés.

2. http://huffingtonpost.com

Version	langue	#articles	longueur	#mots
US	anglais	12 067	454.4	5 482 661
FR	français	4 133	369.6	1 527 416
BR	portugais	2 355	429.5	1 011 373

FIG. 1 – *Brève description des trois corpus constitués pour notre analyse. #articles donne le nombre d'articles publiés, leur longueur moyenne et #mots le nombre total de mots dans le corpus brut (c'est-à-dire non pré-traité, voir section 3.2).*

3.2 Protocole expérimental

Pour le moment, l'approche est divisée en deux étapes : une étape automatique, dans laquelle on extrait les thématiques pour chacun des corpus, et une étape manuelle qui consiste à étudier les thématiques pour leur associer une catégorie transversale à tous les corpus avant de les analyser. Nous discutons dans la section 5 le projet d'intégrer davantage ces deux étapes.

Concernant la partie automatique sur l'extraction des thématiques, nous avons utilisé le modèle LDA (Blei et al., 2003) dans son implémentation parallèle disponible via la librairie MALLET [3]. Les hyper-paramètres du modèle α et β, correspondant respectivement à l'à-priori sur les thématiques et sur les documents, sont symétriques et ont été estimés automatiquement à partir des données (Mccallum et al., 2009). Après plusieurs tentatives manuelles, nous avons choisi de fixer le nombre k de thématiques à 100. D'autres valeurs peuvent évidemment s'avérer pertinentes, mais nous pensons que ce n'est pas tellement important dans la mesure où nous effectuons une analyse comparée de plusieurs corpus. Le nombre d'itérations pour l'estimation avec un échantillonnage de Gibb's a été fixé à 2000, comme préconisé par les créateurs de la librairie. L'algorithme a été exécuté sur une version légèrement nettoyée des données : lettres mises en minuscules, suppression de la ponctuation au début et à la fin des mots (on conserve ainsi "qu'il" ou "sang-froid"), suppression des mots n'apparaissant que dans un seul document et suppression des mots-outils habituels.

La partie concernant l'annotation manuelle a été réalisée par trois chercheurs en sociologie et en sciences de l'information et de la communication. Le processus a été réalisé de manière individuelle dans un premier temps (une personne sur le corpus en anglais et en français, les deux autres sur le corpus brésilien) puis de manière collective afin de converger vers un consensus, à la fois sur les catégories choisies et sur l'annotation. A savoir qu'il était possible qu'une catégorie ne soit présente que pour un ou deux corpus. Finalement, le processus collectif a convergé vers la définition de 15 catégories pour coder l'ensemble des corpus.

3.3 Résultats obtenus

Le tableau de la figure 2 présente un extrait de 5 thématiques, sur les 100 thématiques extraites à partir de chacun des trois corpus.

Nous constatons tout d'abord qu'une majeure partie des thématiques a été conservée car jugée suffisamment pertinente par les sociologues. Plus précisément : 15 ont été écartées en français, 16 en anglais et 16 en portugais. En effet, la plupart des thématiques peuvent être

3. http://mallet.cs.umass.edu/

topic	#doc	cat.	mots les plus probables
			en français (sur 4133 articles) :
z_{18}	28	1	manifestation, paris, police, travail, loi, contre, syndicats, place, bastille, 2016
z_{19}	36	1	loi, travail, gouvernement, l'état, texte, l'assemblée, d'urgence, mois, projet, conseil
z_{25}	39	2	jeux, rio, olympiques, olympique, août, jo, athlètes, 2016, brésil, cérémonie
z_{47}	18	3	morandini, jean-marc, inrocks, catherine, l'animateur, lui, qu'il, europe, comédiens, plainte
z_{73}	47	4	nice, 14, l'attentat, anglais, promenade, camion, attentat, police, soir, christian
			en anglais (sur 12067 articles) :
z_{14}	92	5	refugees, children, refugee, people, countries, world, syrian, rights, million, year
z_{21}	74	2	gymnastics, biles, olympic, team, simone, olympics, gymnast, gold, rio, hernandez
z_{3}	46	6	pokemon, game, pokémon, playing, players, catch, «pokemon, go», pizza, play
z_{50}	56	7	muslim, religious, muslims, faith, church, god, christian, religion, hate, american
z_{27}	140	8	clinton, voters, trump, poll, polls, americans, election, support, vote, relationships
			en portugais (sur 2355 articles) :
z_{44}	52	8	dilma, presidente, impeachment, senado, senadores, processo, senador, rousseff, julgamento, defesa
z_{58}	7	9	sexo, menstruação, durante, rao, mccane, comédia, realmente, corpo, riso, menstruada
z_{71}	11	7	negros, brancos, negras, pessoas, racial, negra, racismo, país, movimento, black
z_{37}	57	2	brasil, vôlei, jogo, medalha, vitória, ouro, seleção, set, brasileiras, torcida
z_{99}	20	7	lgbt, gay, preconceito, violência, sexual, direitos, família, orgulho, estupro, aborto

FIG. 2 – *Extrait des thématiques trouvées dans les trois langues. Les catégories attribuées ici (cat.) correspondent à : 1- Economie / Social, 2- Sport / JO, 3- Show business / people, 4- Sécurité / attentats, 5- Politique étrangère, 6- Technologie / science, 7- Société, 8- Politique nationale, 9- Santé. Les 15 catégories sont visibles dans le graphique de la figure 3.*

interprétées aisément sur la base des premiers mots clefs, comme illustré dans le tableau (nous avons omis les scores pour des raisons de lisibilité). Ce résultat n'est pas surprenant en lui-même car les données en entrée sont de bonne qualité, si on les compare par exemple à des tweets. Il s'agit d'articles de presse souvent bien écrits et d'une longueur suffisante pour en déduire des régularités statistiques. Un examen plus attentif des thématiques rejetées montre cependant qu'une partie d'entre elles (7 dans le cas français) pourrait en réalité être conservées en retournant consulter les données pour faciliter leur interprétation. Dans notre cas, la thématique relative à l'explosion ayant eu lieu à l'aéroport Atatürk en Turquie a ainsi été écartée peut-être un peu rapidement. Nous constatons également une certaine diversité dans les thématiques qui traitent parfois de sujets très génériques (la famille, le cancer ou les images au sens de photographie) ou d'événements très précis (l'attentat de Nice ou la rumeur au sujet du dernier Iphone). Certains sujets traités de manière prépondérante (par exemple les Jeux Olympiques) se retrouvent découpés en plusieurs thématiques (dans notre exemple, une médaille d'or pour la natation, les jeux paralympiques, etc.), ce qui pose des questions sur le caractère homogène du niveau de granularité de l'information analysée.

Ensuite, nous pouvons calculer la distribution de chacun des corpus sur les catégories mises en place. La figure 3 donne une représentation graphique de cette distribution qui a été normalisée afin de gommer la différence dans le volume des articles publiés (trois fois plus d'articles pour la version US que pour la version française, six fois plus que la version brésilienne). Il

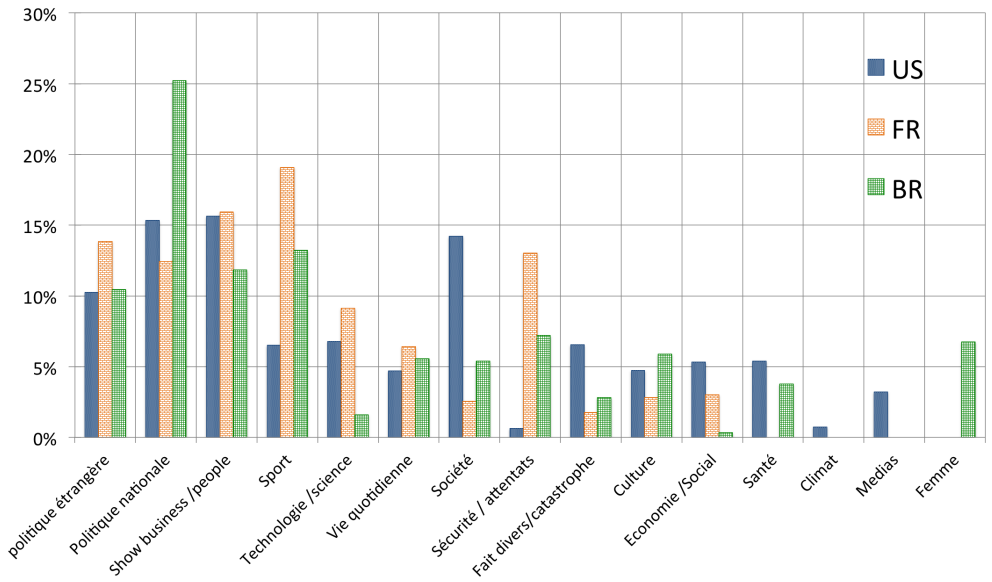


FIG. 3 – *Distribution normalisée des trois éditions sur les 15 catégories, chaque catégorie pouvant être associée à plusieurs thématiques.*

faut rappeler que chaque catégorie est composée de plusieurs thématiques. Par exemple, la politique étrangère est composée de 13 thématiques dans les éditions US et française, alors que le sport est composé de respectivement 12 et 5 thématiques pour ces mêmes éditions.

On peut tirer de ce graphique plusieurs observations préliminaires qui pourront être étudiées plus en détails par la suite. Par exemple, on observe sans grande surprise que la formule mise en place au Huffington Post conduit à mettre en tête la catégorie 'Show business / people' en tête, ou presque si on ne prend pas en compte la catégorie Sport qui est conjecturelle avec la couverture des Jeux Olympiques. L'autre caractéristique de la ligne éditoriale est le spectre très large de la couverture de l'actualité. Sur la politique étrangère, un examen détaillé permet de constater qu'elle est orientée vers les mêmes thématiques : la guerre en Syrie et Irak, les attentats pour la France.

Les écarts structurels entre les deux éditions française et US, qui ne sont que des hypothèses à ce stade de notre travail, sont les suivants. Tout d'abord, il n'existe pas de rubrique Santé dans l'édition française, ainsi que de catégories dédiées au climat ou aux médias, contrairement à l'édition US. Un premier examen des articles associés aux thématiques correspondantes fait ressortir une attention spéciale portée, aux Etats-Unis, aux maladies et, de manière générale, à toutes les manières de "mieux vivre". L'écart constaté sur la rubrique 'Fait divers / catastrophe' s'explique en partie par une place importante des faits divers violents (fusillade, etc.) dans l'actualité américaine. Ces analyses méritent, bien sûr, d'être approfondies et de nombreuses améliorations possibles à court terme sont discutées en section 5.

Comme pour les autres éditions, l'agenda du Huffington Post brésilien respecte l'agenda

médiatique événementiel traditionnel. Les thématiques les plus traitées sont la politique nationale et le sport. La période est marquée par la couverture du processus de destitution de la présidente Dilma Rousseff et des Jeux Olympiques. Au Brésil, la catégorie Show business / people apparaît aussi en tête de liste. La programmation télévisuelle et ses célébrités s'avèrent la principale source de *softnews* du Huffington Post. Notons aussi la place donnée aux problématiques autour du genre (catégorie "femme", mais également "société"), marquées par un discours de dénonciation des violences et infractions des droits.

Cette indexation supervisée par les chercheurs en SHS met en évidence l'arbitraire de toute démarche de catégorisation. Ainsi, les rubriques 'Vie quotidienne' et 'Société' qui cherchent à discriminer délibérément "monde vécu" et "espace public", en suivant Habermas (1987), représentent *de facto* des catégories par défaut, générées par le regroupement d'événements ou de situations échappant à des classifications socio-politiques stables (politique, économie, étranger, etc.). Ces deux catégories génériques sont à développer à moyen terme pour les distribuer en nouvelles thématiques en recourant à des allers-retours entre le niveau macro et les articles eux-mêmes et en isolant éventuellement certains marqueurs sémantiques.

4 Travaux connexes

4.1 Analyse de corpus médiatiques

Les recherches portant sur l'analyse du discours des médias confrontent les chercheurs en sciences humaines à un certain nombre de problèmes méthodologiques. Ils s'accordent sur le fait que la question de la définition d'un évènement ou d'un *thème* évènement pose d'emblée le problème de leur labilité et de leur volatilité ; complexité à laquelle il faut bien ajouter tous les phénomènes de sérialisation qui constituent une des principales caractéristiques de la communication de masse.

Tout corpus qui se voudrait un tant soit peu représentatif prend rapidement l'apparence d'une nébuleuse d'occurrences qu'il s'avère tout simplement difficile de circonscrire. Si l'on écarte une attitude volontariste qui viserait à construire ce dernier en fonction d'hypothèses externes suffisamment fortes, comme c'est le cas pour certaines controverses (Chateauraynaud, 2011), il reste à scruter ce qui demeure observable et tenter alors d'extraire de ce flux des données objectivables. Et, pour ce faire, s'il n'en existe pas, d'envisager l'élaboration de nouveaux instruments d'observation et d'assumer l'option d'une interdisciplinarité "focalisée" avec d'autres sciences (Charaudeau, 2010).

C'est cette instrumentalisation de l'observation et l'étayage de l'analyse à partir de données objectivées à travers de vastes corpus représentatifs qui ont constitué un des principes fondateurs de certaines des recherches précédentes dans le cadre de l'Inathèque de France portant sur les émissions de paroles ou le discours de l'information (Soulages, 2015; Soulages et Lochard, 2016). Ces travaux ont donné lieu à la création de base de données relationnelles dédiées à différents corpus de productions médiatiques et à l'élaboration d'un outil informatique destiné aux chercheurs en SHS (Médiacorpus).

Cette démarche, au départ non supervisée qui repose sur le traitement automatisé de flux de news, met au jour des corrélations ou des facteurs récurrents invisibles à l'observation non appareillée de l'analyste de discours. De plus, des aller retour fréquents entre les résultats produits par l'indexation automatique des données et des séquences d'analyse contextualisées

et compréhensives à l'initiative du chercheur en SHS, visent à la fois à améliorer l'efficacité de l'outil de fouille des données mais aussi à affiner l'analyse qualitative des discours médiatiques étudiés.

4.2 Agrégation par modèles thématiques

L'analyse comparée de plusieurs corpus médiatiques a donné lieu à de nombreux travaux en SHS. Lorsqu'on cherche à faire une analyse un tant soit peu quantitative, celle-ci nécessite l'annotation manuelle de nombreuses ressources (Powers et Benson, 2014). L'utilisation de thématiques peut être vue comme un *proxy* permettant de faciliter cette comparaison. En effet, il est clairement moins coûteux d'annoter un nombre limité de thématiques en lieu et place de l'intégralité du corpus, voire même un sous-ensemble de ce corpus qui nous ramènerait à l'apprentissage supervisé dont nous avons déjà souligné les limitations. Ces techniques automatiques ont déjà été testées dans la pratique du journalisme (Rusch et al., 2013; Günther et Quandt, 2016) mais, à notre connaissance, jamais dans une perspective inter-corpus, qui plus est rédigés en plusieurs langues.

L'extraction des thématiques est associée à une littérature importante en informatique depuis le projet *Topic Detection and Tracking* initié par Allan et al. (1998). Elle peut se baser sur des méthodes de clustering géométrique similaires aux k-moyennes (Velcin et Ganascia, 2007), mais plus souvent sur des factorisations de matrices, comme dans le cas de LSA (Deerwester et al., 1990) ou NMF (Paatero et Tapper, 1994), et sur des approches probabilistes, avec des modèles comme pLSA (Hofmann, 1999) ou LDA (Blei et al., 2003). Bien que voisines des techniques de clustering plus traditionnelles qui attribuent une unique catégorie à un texte, il ne faut pas confondre les deux tâches (Velcin et al., 2016). Lorsque les textes sont suffisamment longs, comme c'est le cas ici avec les articles de presse, il nous semble important de pouvoir associer plusieurs thématiques à un même texte. Cela ouvre également des perspectives très intéressantes pour étudier les liens existants entre les thématiques, et ainsi estimer une structure entre celles-ci et améliorer la compréhension du corpus.

Dans notre travail, nous avons choisi le modèle LDA pour sa popularité et sur la base de ses bons résultats obtenus dans de nombreuses applications (Hall et al., 2008), bien que d'autres modèles pourraient être choisis sans remettre en cause l'approche que nous proposons. De plus, des variantes permettent de prendre en compte des dimensions additionnelles, telles que les auteurs (Rosen-Zvi et al., 2004) ou le temps (Wang et McCallum, 2006). Cela ouvre donc de nombreuses perspectives pour prendre en compte ces dimensions complémentaires dans notre analyse. Malgré cela, nous souhaitons mettre au point une démarche qui ne dépend pas de la nature du modèle employé, que ce dernier soit basé sur des similarités, de l'algèbre linéaire ou des modèles probabilistes.

5 Conclusion et perspectives

Dans cet article, nous présentons une première analyse comparée des trois éditions différentes d'un média international en ligne. Les travaux étant encore dans leur première phase, nous avons mis l'accent sur la méthodologie déployée et donné uniquement quelques résultats préliminaires qui demandent à être approfondis. On constate cependant clairement plusieurs éléments saillants, telles que la couverture de nombreuses thématiques quelque soit l'édition,

les catégories communes (la politique, le show business), mais aussi les spécificités locales (les attentats ou le sport en France, la santé ou les thèmes de société aux USA, les questions de genre au Brésil). Ce travail nous a permis de mettre au jour plusieurs points faibles qui sont autant de perspectives à court terme que nous envisageons d'explorer.

Tout d'abord, l'élaboration des catégories reste un processus manuel qui ne peut être totalement traité comme une tâche de classification supervisée. En effet, il nous paraît évident qu'il est risqué de se reposer sur les catégories à priori proposées sur le site du (ou des) média(s), en particulier dans une optique diachronique. Pour un même média, ces catégories changent régulièrement en fonction de l'actualité, mises à part peut-être certaines d'entre elle, et celles-ci sont rarement les mêmes d'un média à l'autre. D'un autre côté, il est nécessaire d'avoir un socle commun pour réaliser une comparaison et l'opération ne peut pas être totalement réalisée sur la base des thématiques. Les modèles proposés par Paul et Girju (2009) ou Chen et al. (2015) vont dans ce sens mais ils ne sont pas aisément extensibles et, pour le moment, dédiés à une seule langue. Une approche semi-supervisée, dans laquelle il serait possible de combiner des catégories à priori avec des catégories émergentes semblerait plus appropriée, dans un esprit intégré discuté par Chuang et al. (2014). La prise en compte d'à priori provenant des experts sous la forme de collections de mots-clefs, comme dans les travaux de Newman et al. (2011) ou Hu et al. (2014), paraît une bonne piste pour cela. Nous envisageons également de recourir à des techniques d'alignement entre thématiques, rendues plus difficiles ici du fait de la diversité des langues employées. Bien que les résultats sur trois éditions aient été présentés ici, nous visons toujours l'intégration de l'édition arabe du Huffington Post. Contrairement aux travaux réalisés par Mimno et al. (2009), nous n'avons pas ici les mêmes articles accessibles dans plusieurs langues et une plus grande hétérogénéité en fonction des sources.

Ensuite, une autre perspective qui nous paraît importante consiste à aider l'expert à comprendre le sens des thématiques, ce qui lui permet ensuite d'attribuer plus facilement des catégories et de réaliser ainsi son analyse. C'est pourquoi nous travaillons actuellement à l'intégration de techniques de nommage des thématiques (*topic labeling*, voir Mei et al. (2007)). Cet étiquetage automatique des catégories thématiques, qu'il opère au niveau des thématiques initiales ou des grandes catégories, peut reposer par exemple sur des n-grams. Ainsi, la thématique z_{18} (France) du tableau 2 peut assez facilement être étiquetée par les termes "loi travail" et/ou "manifestation contre la loi". De la même façon, la thématique z_3 (US) peut être étiquetée par "augmented reality game" et/ou "pokemon go game". Au-delà d'un titre compréhensible (Lopez et al., 2014), différentes techniques de recherche d'information et de résumé automatique peuvent être mobilisées pour donner à l'expert une bien meilleure compréhension de la cohérence issue des régularités statistiques.

Enfin, nous prévoyons d'automatiser l'identification des thématiques non pertinentes en ayant recours aux indices de qualité développés dans la littérature (Röder et al., 2015). Nous avons également comme objectif de comparer l'information issue des articles de presse de celle publiée sur les billets de blogs hébergés par le Huffington Post, mais également d'intégrer l'édition arabe et d'étendre l'analyse pour prendre en compte la dimension temporelle (comment les thématiques ou les catégories évoluent-elles dans le temps ?).

Références

Allan, J., J. G. Carbonell, G. Doddington, J. Yamron, et Y. Yang (1998). *Topic detection and tracking pilot study final report*.

Appadurai, A. (2011). Disjuncture and difference in the global cultural economy 1990. *Cultural theory : An anthology 2011*, 282–295.

Aubrun, F. (2015). *Crise(s), publicité et marque : L'émergence de nouveaux modèles*. Ph. D. thesis, Université Lumière Lyon 2.

Blei, D. M., A. Y. Ng, et M. I. Jordan (2003). Latent dirichlet allocation. *Journal of machine Learning research 3*, 993–1022.

Charaudeau, P. (2010). Pour une interdisciplinarité "focalisée" dans les sciences humaines et sociales. *Questions de communication* (17), 195–222.

Charon, J.-M. et J. Papet (2014). *Le journalisme en questions : réponses internationales*. Paris : L'Harmattan.

Chateauraynaud, F. (2011). *Argumenter dans un champ de forces. Essai de balistique sociologique*. Paris : éditions Petra.

Chen, C., W. Buntine, N. Ding, L. Xie, et L. Du (2015). Differential topic models. *IEEE transactions on pattern analysis and machine intelligence 37*(2), 230–242.

Chuang, J., J. D. Wilkerson, R. Weiss, et al. (2014). Computer-assisted content analysis : Topic models for exploring multiple subjective interpretations. In *Advances in Neural Information Processing Systems Workshop on Human-Propelled Machine Learning*.

Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, et R. A. Harshman (1990). Indexing by latent semantic analysis. *JASIS 41*(6), 391–407.

Günther, E. et T. Quandt (2016). Word counts and topic models : Automated text analysis methods for digital journalism research. *Digital Journalism 4*(1), 75–88.

Habermas, J. (1987). *Théorie de l'agir communicationnel*. Paris : Fayard.

Hall, D., D. Jurafsky, et C. D. Manning (2008). Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pp. 363–371. Association for Computational Linguistics.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the 15th conference on Uncertainty in Artificial Intelligence*, pp. 289–296. Morgan Kaufmann.

Hu, Y., J. Boyd-Graber, B. Satinoff, et A. Smith (2014). Interactive topic modeling. *Machine learning 95*(3), 423–469.

Lopez, C., V. Prince, et M. Roche (2014). How can catchy titles be generated without loss of informativeness ? *Expert Systems With Applications (ESWA) 41*(4), 1051–1062.

Mccallum, A., D. M. Mimno, et H. M. Wallach (2009). Rethinking lda : why priors matter. In *Advances in Neural Information Processing Systems*, pp. 1973–1981.

Mei, Q., X. Shen, et C. Zhai (2007). Automatic labeling of multinomial topic models. In *ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pp. 490–499.

Mimno, D., H. M. Wallach, J. Naradowsky, D. A. Smith, et A. McCallum (2009). Polylingual topic models. In *Proceedings of the ACL Conference on Empirical Methods in Natural*

Language Processing : Volume 2, pp. 880–889.

Newman, D., E. V. Bonilla, et W. Buntine (2011). Improving topic coherence with regularized topic models. In *Advances in neural information processing systems*, pp. 496–504.

Paatero, P. et U. Tapper (1994). Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics 5*(2).

Paul, M. et R. Girju (2009). Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the ACL conference on Empirical Methods in Natural Language Processing : Volume 3*, pp. 1408–1417.

Powers, M. et R. Benson (2014). Is the internet homogenizing or diversifying the news ? External pluralism in the US, Danish, and French press. *The International Journal of Press/Politics 19*(2), 246–265.

Röder, M., A. Both, et A. Hinneburg (2015). Exploring the space of topic coherence measures. In *Proceedings of the ACM international conference on Web Search and Data Mining*.

Rosen-Zvi, M., T. Griffiths, M. Steyvers, et P. Smyth (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 487–494. AUAI Press.

Rusch, T., P. Hofmarcher, R. Hatzinger, K. Hornik, et al. (2013). Model trees with topic model preprocessing : An approach for data journalism illustrated with the wikileaks afghanistan war logs. *The Annals of Applied Statistics 7*(2), 613–639.

Soulages, J.-C. (2015). Le traitement télévisuel du sida en france : apprivoisement du fléau, instrumentation des médias. In I. Pânzaru et D. Popescu-Jourdy (Eds.), *Nouvelles approches de la rationalité. Défis contemporains des sciences humaines et sociales*, pp. 229–257. PUL.

Soulages, J.-C. et G. Lochard (2016). Comment la télévision traite la laïcité. In *La laïcité dans l'arène médiatique. Cartographie d'une controverse sociale*, pp. 95–115. INA éditions.

Velcin, J. et J.-G. Ganascia (2007). Topic extraction with AGAPE. In *Advanced Data Mining and Applications*, pp. 377–388. Springer.

Velcin, J., M. Roche, et P. Poncelet (2016). Shallow text clustering does not mean weak topics : how topic identification can leverage bigram feature. In *Proceedings of DMNLP, Workshop at ECML/PKDD*, Riva del Garda, Italy, pp. 25–32.

Wang, X. et A. McCallum (2006). Topics over time : a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 424–433. ACM.

Summary

This article presents a joint work undertaken by a group of researchers in computer science, sociology and communication and information sciences, in the context of the digital journalism initiative. In the design of a new analysis process, we build a manual encoding upon a first automatic, unsupervised topic extraction from textual data provided on the Huffington Post website. This approach makes possible a comparative study of news published during the summer 2016 in three editions of the journal (French, English, Brazilian). The first presented results validate the whole process but gives room for improving many steps of the process.

Génération automatique de billets journalistiques : singularité et normalité d'une sélection

Jérémy Vizzini, Cyril Labbé, François Portet

Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
CNRS, LIG, F-38000 Grenoble, France
nom.prenom@imag.fr

Résumé. La génération automatique de billets journalistiques est utilisée pour transcrire en texte des informations répondant à une requête utilisateur : prévisions météo, résultats d'élections, sportifs ou boursiers... En général, le texte généré se contente de décrire les données pertinentes sans souligner leurs singularités ou normalité par rapport à un sur-ensemble. Or, la plupart du temps, les connaissances disponibles permettent de souligner ce type d'informations. En effet, les données décrivent une même information à différents instants ou pour différents groupes. Il est donc envisageable de détecter automatiquement le caractère exceptionnel d'un prix ou d'une défaite électorale. Cet article présente une démarche et un outil permettant de spécifier un générateur de texte comparant une sélection de données à l'ensemble des données disponibles. Les singularités et les régularités exprimables à partir d'une sélection de données sont illustrées à travers un exemple (résultats d'élections). L'explicitation des connaissances (modèles, ressources langagières,...) nécessaires à la construction du générateur de textes a pour objectif de rendre l'approche générique et applicable à différents domaines (météo, élections, sports,...). Le prototype `Summy` a été réalisé pour valider la faisabilité de l'approche en utilisant les données des élections régionales. Ce prototype permet notamment de montrer comment automatiquement exprimer en langage naturel la singularité ou la normalité d'un sous-ensemble de données par rapport à l'ensemble.

1 Introduction

Si on dit qu'un graphique vaut mieux qu'un long discours il peut être aussi avancé qu'un petit texte synthétique, ciblé et nuancé vaut mieux que de nombreux graphiques, ou du moins qu'il vient les compléter avantageusement. En effet, la richesse des langues permet de mieux résumer des informations complexes et nombreuses en les rendant accessibles à tous : un texte peut être écouté par des malvoyants et le discours peut être adapté à l'expertise du destinataire. Le domaine de la génération automatique de textes (GAT) offre donc des perspectives pertinentes et intéressantes pour transmettre des connaissances riches, complexes et personnalisées.

Cependant, le développement d'un système de GAT est fortement dépendant du domaine et de l'application ciblée. De plus, si pour la génération à partir de données (data to text), malgré

la qualité indéniable des textes générés, on peut observer que ceux-ci restent dans une visée purement descriptive des données sélectionnées. Or, lors de la communication d'une information, il est souvent intéressant de contextualiser le contenu avec des informations exprimant des similarités et différences observables. On peut ainsi mentionner des évolutions ou des corrélations en rapport avec la sélection mais n'en faisant pas partie explicitement. Par exemple, les prévisions météo ou les résultats d'élection concernant une localité peuvent être comparés aux données concernant des localités ayant des propriétés identiques, p.ex., de la même région etc. Comparaisons qui peuvent être ensuite être insérées dans le texte.

Cet article présente une approche pour résumer des données sous forme textuelle tout en mettant en exergue les singularités et/ou la normalité d'un sous-ensemble peuvant être transcrites dans un texte. La démarche consiste à identifier et expliciter les ressources et connaissances (modèles, ressources langagières etc.) nécessaires à la construction du générateur de texte. L'objectif est de rendre l'approche générique et applicable à moindre coût dans différents domaines (météo, élections, sports...). Un prototype `Summy` a été réalisé et testé avec des données d'élection régionales.

La section 2 présente un exemple concret d'application : la génération de résultats d'élection qui sera repris tout le long de l'article ; qui permettra de mettre en lumière l'intérêt de l'approche. La section 3 décrit le processus de génération, l'architecture fonctionnelle et les différents traitements. La section 4 présente plus en détail chacune des phases de traitement en formalisant la notion de sélection de données et en décrivant le processus de macro et micro planification de textes. La section 4 présente le prototype et un bref positionnement par rapport à la littérature du domaine (section 5) avant de conclure en section 6.

2 Enrichir les textes décrivant les résultats des élections

La Figure 1 présente une modélisation possible des données représentant les élections régionales françaises de 2015. Les résultats sont disponibles sur `data.gouv.fr` pour les 18 régions françaises ainsi que pour chaque département et commune (data.gouv.fr, Plateforme ouverte des données publiques françaises, 2016). Pour les sites d'information, le défi réside en la génération en un temps minimal d'un grand nombre de textes présentant les résultats individualisés de chaque zone de vote. Étant donné que pour des élections de grande ampleur il serait bien trop coûteux de faire réaliser le travail de rédaction par l'homme, certains sites d'information ont mis en place des systèmes de génération automatique de textes tels que *The Associated Press* avec la solution d'`Automated Insights`, ou encore le site `lemonde.fr` qui a utilisé la solution `Data2Content` développée par la société (Syllabs, 2016).

L'exemple 1 est un exemple de texte généré automatiquement et publié sur le site `lemonde.fr`. Le texte est composé en premier lieu d'un préambule où les candidats de la région sont présentés. Ensuite, les résultats de la commune sont donnés. Ce texte est fluide et conforme aux attentes de l'utilisateur. Cependant, il contient exactement et uniquement les informations de la ville de Grenoble en 2015 au deuxième tour [1].

Or la base de données contient des informations d'autres temporalités (les élections précédentes) ainsi que les résultats des autres villes, des autres départements et régions. Techniquement, il est donc possible, à l'aide de ces connaissances, de générer un texte contenant d'autres

1. L'absence d'autres informations peut provenir de différentes causes : difficultés techniques, risque d'erreurs plus important ou simplement problèmes d'acceptabilité d'un point de vue éthique, sociale ou déontologique.

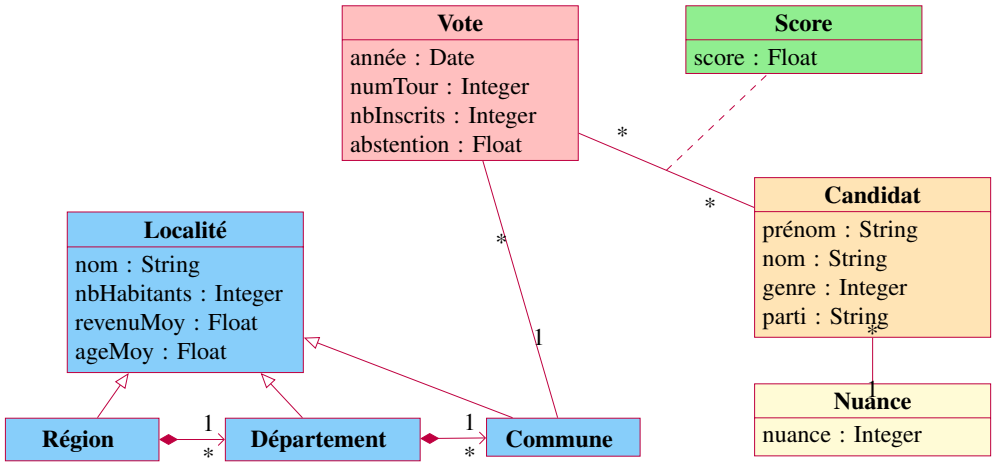


FIG. 1 – *Modélisation des données des élections régionales*

Exemple 1 (Extrait du texte des résultats de la ville de Grenoble (`Data2content`).). *Au second tour des élections régionales 2015, la région Auvergne-Rhône-Alpes voyait trois candidats s'affronter : M. Jean-Jack Queyranne (liste Union de la Gauche), M. Laurent Wauquiez (liste Union de la Droite) et M. Christophe Boudot (liste Front National).*
Voici les résultats du second tour pour la ville de Grenoble : la liste Union de la Gauche de M. Jean-Jack Queyranne est arrivée en première position avec 57,19 % des voix. Elle a devancé la liste Union de la Droite de M. Laurent Wauquiez qui a obtenu 29,78 % et la liste Front National de M. Christophe Boudot , qui a recueilli 13,03 % des voix. Dans cette localité, 46,51 % des inscrits se sont abstenus.

informations contextuelles. Pour cela, il faut estimer le niveau de similitude entre les données sélectionnées par l'utilisateur (résultats dans la commune) et d'autres ayant une caractéristique similaire (dates différentes, autres communes de la région, communes de taille similaire...). Lors de la génération du texte, il est intéressant de souligner le caractère *normal* ou *exceptionnel* de ces élections. Par exemple, souligner que le parti du gagnant dans la commune est commun aux autres villes de la région ou au contraire exceptionnel, signaler que le score du parti vainqueur est habituel ou non par rapport à l'historique de la ville. Dans les deux cas, l'ensemble sélectionné (résultats dans la commune) est comparé à un sur-ensemble composé d'éléments ayant une caractéristique commune (villes différentes, même région ou même ville, dates différentes).

L'exemple 2 concerne le même sujet que l'exemple 1 mais présente plus d'information contextuelle. Le lecteur obtient à la fois les informations sélectionnées mais aussi leurs mises en perspective qui permet une interprétation plus large.

Cet article présente une démarche et un outil permettant de construire un générateur de texte personnalisable et offrant la possibilité d'inclure les enrichissements présentés ci-dessus. Dans la suite de l'article, seule la singularité ou la régularité du taux d'absentions servira d'exemple. La démarche se veut générique et adaptable à de nombreux domaines.

Exemple 2 (texte des résultats des élections avec enrichissements)**.** *Au second tour des élections régionales 2015, la région Auvergne-Rhône-Alpes voyait trois candidats s'affronter : M. Jean-Jack Queyranne (liste Union de la Gauche), M. Laurent Wauquiez (liste Union de la Droite) et M. Christophe Boudot (liste Front National).*
Les résultats de la ville de Grenoble sont très similaires *à ceux des autres communes de la région. La liste Union de la Gauche est arrivée en première position avec 57,19 % des voix. Elle a devancé la liste Union de la Droite qui a obtenu 29,78 % et la liste Front National, qui a recueilli 13,03 % des voix.*
Cette tendance à voter à Gauche est habituelle *pour une ville ayant une population jeune aux revenus moyens . Entre le premier et le second tour, le taux d'abstention a* diminué *de près de 10 %, passant de 55,50 % à 46,51 %. Un tel score et une telle diminution ne sont pas* surprenants *au regard des précédentes élections régionales.*

3 Processus de génération

Un concepteur de générateur de textes, spécialiste de son domaine (élection, météo, sport) doit pouvoir définir chacune des étapes de la génération et créer ou compléter les ressources et connaissances du domaine nécessaires à l'interprétation des données. La section 3.1 présente le processus global de génération de textes enrichis ainsi que les ressources nécessaires. Les sections suivantes détaillent chacune de ces étapes.

3.1 Vue d'ensemble

L'approche que nous avons mise en œuvre s'inspire de l'architecture en pipeline du domaine de la GAT telle que présentée par (Reiter et Dale, 2000). La génération commence par la détermination du contenu qui ici est soumise à la requête de l'utilisateur (Figure 2 *Sélection du contenu*). Dans notre domaine applicatif, cela correspond à sélectionner une "entité d'intérêt" à décrire parmi l'ensemble des données disponibles. Par exemple, pour le cas des élections, le lecteur sélectionne la ou les villes dont il veut connaître les résultats. Par ailleurs, des fonctions d'évaluation et de comparaison permettent d'obtenir des informations provenant du contenu sélectionné (nombre de tuples, max, min,...) et du reste de l'ensemble. Ces fonctions de comparaison sont définies par le concepteur : elles mesurent la similarité ou dissimilarité de la sélection par rapport à un autre ensemble.

Puis la structure du document est déterminée (Figure 2 structuration du document). Cette structure est spécifiée par le concepteur à l'aide d'un langage impératif de haut niveau décrivant les sections, les paragraphes et les phrases du document.

Le processus se termine par le micro-planning (Figure 2, finalisation des phrases et réalisation de surface). Cette phase s'appuie sur les structures de phrases et le lexique nécessaire à la génération en langage naturel des données de la base. Ces connaissances se composent d'un ensemble de dictionnaires associés à la base. Ces dictionnaires peuvent être préexistants ou enrichis par le concepteur.

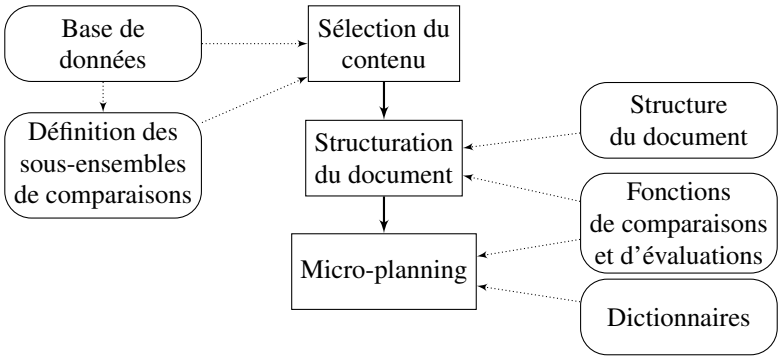


FIG. 2 – *Schéma du processus de génération*

3.2 Sélection du contenu

Quand l'entité sur laquelle le résumé doit porter est sélectionnée (dans l'exemple, la ou les villes), `Summy` utilise deux ensembles de données pour générer le texte décrivant l'entité et les comparaisons.

Le premier ensemble de données est qualifié de "primaires" (TP dans la suite), il regroupe toutes les connaissances disponibles sur l'entité sélectionnée par l'utilisateur. TP doit être défini par le concepteur du générateur de textes.

Exemple 3. *Le lecteur choisit un ensemble de villes. L'ensemble de données primaires peut donc être une table regroupant l'ensemble des informations reliées à la ville (cf. figure 1) :* $TP_VILLES(Annee, Num_Tour, Nom_Departement, Nom_Region, Nom_Commune, Habitants, Revenu_Moy, Age_Moy, Inscrits, Abstention)$

Le deuxième ensemble de données est dit "secondaire" (TS), il regroupe l'ensemble des informations qui seront résumées textuellement.

Exemple 4. *Dans l'exemple des élections,* TS *est constitué des scores de chaque candidat dans chaque localité :* $TS_CANDIDATS(ID_Localite, Annee, Num_Tour, Annee, Prenom, Nom, Genre, Parti, Nuance, Score)$

De ces deux ensembles on en déduit S qui est l'ensemble des données qui serviront à la génération du texte final. S est une restriction (au sens de (Codd, 1970)) de l'ensemble TS par l'ensemble des données choisi par l'utilisateur dans TP. La restriction est définie par un ensemble de propriétés communes entre les deux ensembles TP et TS. Dans la suite, cet ensemble de propriétés communes sera nommé "filtre".

Exemple 5. *Pour l'exemple des élections, la sélection* S *des données à transcrire est un sous-ensemble de* `TS_CANDIDATS` *où les tuples concernent la même commune, la même année et le même tour donnant ainsi la liste des candidats pour une ville sélectionnée dans l'ensemble primaire* `TP_VILLES`.

Les ensembles de données ainsi que les filtres doivent être définis par le concepteur. Les filtres sont aussi utilisés par le concepteur pour définir les ensembles auxquels S doit être

comparé. Par exemple, l'ensemble des résultats pour les villes de même région que la ville de S.

Exemple 6. *La comparaison de l'abstention dans une ville peut se faire par une restriction de la table TP et le filtre $\{Nom_Region\}$.*

3.3 Structuration du document : macro planning

`Summy` propose un langage de haut-niveau s'inspirant des schémas (McKeown, 1985) permettant de placer les éléments structurants du texte : les sections, les paragraphes et les listes.

Le contexte courant du générateur est défini par un ensemble de données (i.e. une table), un tuple courant et un ensemble de variables globales contenant des métadonnées (nombre de tuples de la table courante, résultat de comparaison,..). Une liste de directives et de variables globales permettant de définir la structure d'un texte sont données dans l'exemple 7.

Exemple 7.

- `Iteration(table_name)` *itération des tuples de la table* `table_name`*, fixe la table courante.*
- `Sentence(sent_id)` *ajoute au texte la phrase identifiée par* `sent_id`*. La phrase est construite en utilisant le contexte courant (table, tuple , variables globales).*
- `Link(link_name)` *ajoute au document un mot de liaison* `link_name`*.*
- `Similarity(evaluator, filter_id)` *mesure de la* normalité *du tuple courant à l'aide de* `evaluator`*. Le filtre d'identifiant* `filter_id)` *définie la restriction de la table courante par le tuple courant.*
- `Paragraph()` *ajoute un nouveau paragraphe.*
- `COUNT` *variable globale contenant le nombre de tuples de la table courante.*
- `SIMILARITY` *variable globale contenant le score de 0 à 1 de la dernière fonction de similarité calculée.*
- `FILTER` *variable globale contenant l'identifiant du dernier filtre utilisé.*

La figure 3 donne un exemple de structuration de document. Le document commence par une phrase préambule identifiée par `phrase_preambule`. Puis on énumère (`Iteration()`) les villes sélectionnés par le lecteur. Pour chaque ville, les candidats sont générés avec leur score électoral. Enfin, une recherche de similarité est effectuée entre le taux d'abstention du tuple courant (`TU_GET(ABSTENTION)`) par rapport à un ensemble défini par le filtre (`filtre_meme_tour_region`). Le comparateur $Similarity(evaluator, filter)$ est à définir par le concepteur en fonction de la sémantique associée aux données. Différents algorithmes de fouille de données et/ou de classification peuvent être utilisés. Pour l'abstention, une valeur est considérée exceptionnelle ou normale en fonction de son écart à la moyenne et de l'écart-type observé dans l'ensemble de comparaison.

3.4 Lexicalisation des variables

Deux types différents d'éléments textuels peuvent être placés dans le document : les mots de liaisons et les modèles de phrases.

```
1   Sentence(phrase_preambule)
2
3   Iteration() {
4       Paragraph()
5       Sentence(phrase_candidats_region)
6
7       Sentence(phrase_candidats_score)
8       Iteration(PROJ_CANDIDATS) {
9           Sentence(phrase_score_parti)
10      }
11
12      Similarity(TU_GET(ABSTENTION), filtre_meme_tour_region)
13      Sentence(phrase_abstention_vile)
14  }
```

FIG. 3 – *Exemple de macro-planning*

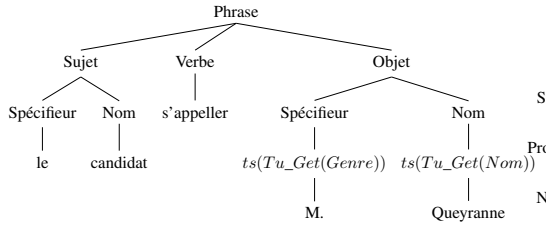


FIG. 4 – *Modèle de phrase contenant deux fonctions de lexicalisation.*

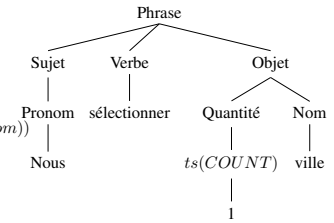


FIG. 5 – *Modèle de phrase contenant une fonction de lexicalisation.*

Les mots de liaison ont été séparés des modèles de phrases car ceux-ci sont dépendants de la structure du document : par exemple les mots de liaison indiquant des énumérations. Lors d'une énumération des mots tel que 'premièrement', 'deuxièmement', etc peuvent être ajoutés. Au contraire, le système doit supprimer cette liaison si une seule ville est transcrite.

Les phrases sont définies sous une forme permettant la réalisation de surface. La structure de la phrase est donnée sous forme d'arbre dont certaines feuilles doivent être calculées au moment de la lexicalisation qui définit les termes à associer à une donnée particulière (cf. Figure 4 et 5).

La fonction de lexicalisation ts associe un mot à une valeur d'attribut, à un attribut ou à un objet du modèle. `Summy` implémente ces fonctions sous forme de dictionnaires.

La fonction de lexicalisation agit sur une donnée résultant de l'appel à un évaluateur. Un évaluateur est une fonction qui associe une valeur à une ensemble de données. Par exemple, $Tu_Get(Tuple, Attr)$ renvoie la valeur du tuple pour l'attribut donné ou encore $Ta_MostFreq(Table, Attr)$ calcule la valeur la plus fréquente pour l'attribut donnée.

Des modèles génériques sont mis à disposition du concepteur qui doit uniquement complé-

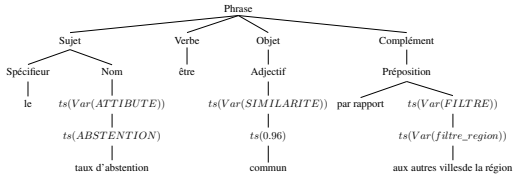


FIG. 6 – *Modèle de phrase générique attaché à la similarité*

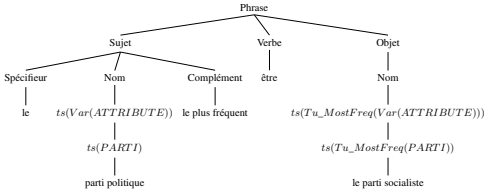


FIG. 7 – *Modèle de phrase générique attaché à $Tu_MostFreq$*

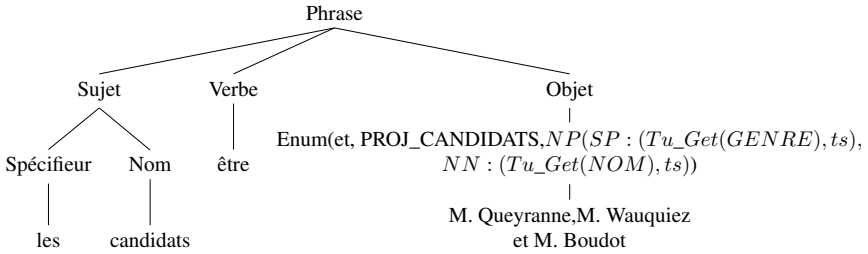


FIG. 8 – *Modèle de phrase contenant une énumération*

ter la génération des noms, des attributs ou le nom des filtres utilisés. Les figures 6 et 7 sont des exemples de modèles de phrases génériques. Ils sont fournis par le système au concepteur qui utilise leur identifiant lors de la structuration du document 3.3.

Un opérateur particulier permet d'exprimer une énumération. La figure 3.4 montre un exemple d'itération sur les différents tuples de $TS_CANDIDAT$. Cet opérateur permet à `Summy` de ne pas mettre en œuvre l'étape d'agrégation consistant à regrouper les phrases décrivant chacun des tuples.

4 Implémentation de Summy

L'implémentation du prototype a été effectuée en `Java` et utilise la version bilingue (franco-anglaise) de la bibliothèque `SimpleNLG` pour la réalisation de surface (Vaudry et Lapalme, 2013). Les tests ont été effectués sur les résultats des élections régionales françaises de 2015 stockés dans un SGBD (`MySQL`) selon le modèle décrit dans la section 3.2. La table `TP_VILLES` comporte 8092 tuples, la table `TS_CANDIDATS` 456699.

Le prototype est composé d'une interface graphique qui permet au concepteur de définir un fichier de configuration (format `JSON`) comportant les ressources linguistiques et informationnelles décrites Figure 2. Le concepteur définit les ensembles de données (TS,TP,S), les filtres et fonctions pour les comparaisons. La structure des documents à générer est définie à l'aide du langage présenté dans la section 3.3.

Lors de la phase de génération, l'application se connecte à la base de données et effectue la génération de texte. Le fichier de configuration ainsi que l'archive `jar` du prototype peuvent être intégrés à des applications afin de mettre en place facilement un système de génération de textes directement au sein d'un logiciel métier. Les dictionnaires ont été créés à partir des textes disponibles sur `lemonde.fr` afin d'y ajouter les enrichissements de manière à générer des textes tels que celui de la figure 2.

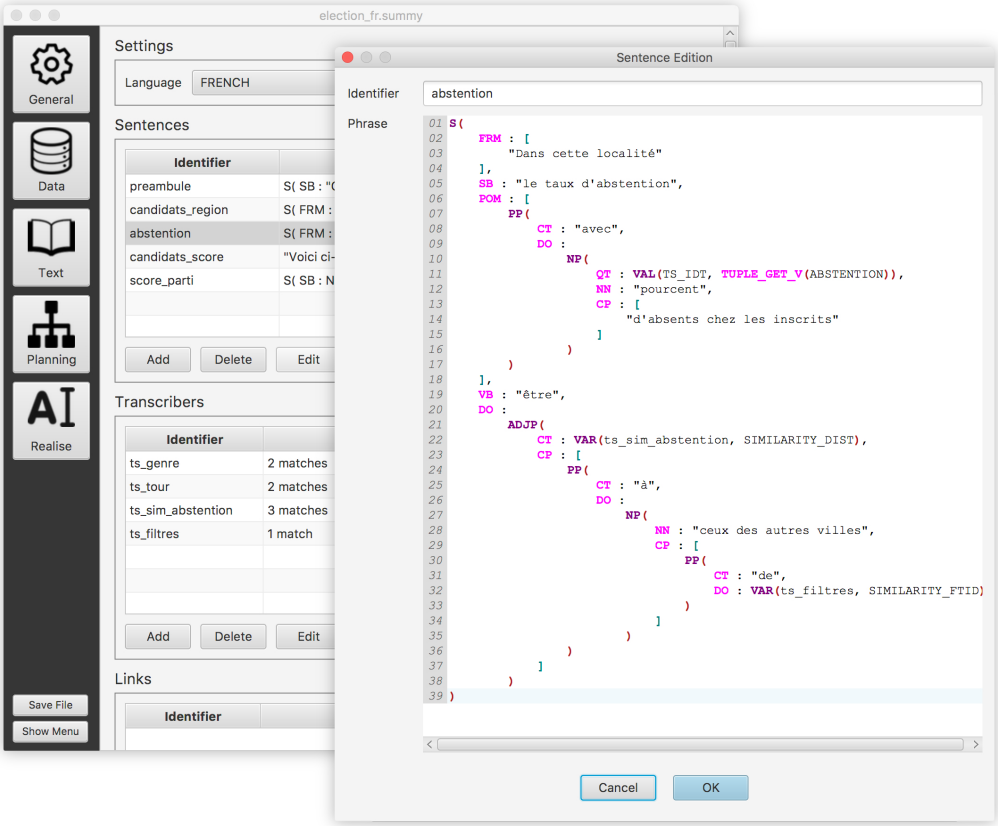


FIG. 9 – *Interface d'édition d'un modèle de phrase*

5 Travaux relatifs

Le résumé textuel d'informations structurées a déjà fait l'objet de nombreux travaux notamment dans le domaine de la Génération Automatique de Textes (GAT), des Bases de Données et du web sémantique. On peut citer ONTOSUM (Bontcheva, 2005) et NaturalOWL (Androutsopoulos et al., 2013) dont l'objectif est de rendre accessible des informations contenues dans une ontologie de type RDF/OWL par sélection d'un concept et génération d'une description

textuelle des propriétés et concepts liés. L'architecture adoptée étant le pipeline classique de la GAT. Si ces approches contribuent de manière significative à un web sémantique plus accessible (cf. le KBGen challenge (Banik et al., 2013)) elles reposent cependant sur une sémantique très riche bien loin des bases de données classiques.

Un autre domaine d'application est celui du résumé textuel de graphiques à destination des personnes ayant des déficiences visuelles (synthèse vocale de descriptions). En effet les systèmes tels de SIGHT (Demir et al., 2012) ou iGraph (Dumontier et al., 2010) permettent à l'aide de primitives de graphiques d'extraire les segments pertinents d'un graphique et de les décrire textuellement afin d'augmenter l'accessibilité des pages web. Si certains éléments de discours sont communs avec notre approche (comparaison) ces approches restent spécifiques au domaine des graphiques.

Le domaine des bases de données est assez riche en solutions de génération de textes à partir de données, que cela soit dans le domaine industriel (Syllabs dont le slogan est "Faites parler vos données", Arria, Yseop, etc.) qui surfe sur le *data analytics* ou dans le domaine académique (Labbé et al., 2015; Portet et al., 2009; Labbé et Portet, 2012). Si dans l'approche développée dans Babytalk (Portet et al., 2009) un ensemble important et hétérogène de données médicales sont analysées, liées par des relations rhétoriques et résumées en texte, cette approche a requis un grand nombre de connaissances du domaine et n'est donc pas facilement portable dans un autre domaine. L'approche développée par (Labbé et al., 2015), présente quant à elle un système capable de générer du texte à partir du résultat de requêtes de base de données. Le langage de ces requêtes permet de restreindre la génération à des primitives génériques pour lesquelles le passage d'un domaine à un autre est trivial. Ce système n'embarque cependant pas de comparaison automatique et reste donc uniquement guidé par les primitives disponibles pour la requête. L'approche présentée dans cet article permet donc d'apporter un complément utile d'un point de vue analyse des données à (Labbé et al., 2015) tout en conservant l'aspect générique du système.

6 Conclusion

Dans cet article, nous avons proposé une approche pour résumer dans un texte des données issues d'une requête utilisateur. Cette approche a été implantée dans un logiciel et testée sur des données de résultats d'élections. Elle permet de concevoir une application de génération de textes avec une paramétrisation réduite. La prochaine étape consistera à mettre en place une évaluation sur un ensemble plus diversifiés de données (sport, bourse, etc.) en mesurant d'une part, les performances techniques et qualitatives (temps d'exécution, qualité linguistique et cohérence du texte généré). Il conviendra en particulier de mesurer le temps de développement nécessaire pour les utilisateurs-développeur et d'évaluer la préférence des lecteurs lors d'une étude comparative. Une réflexion doit aussi être menée sur la pertinence et la validité des faits automatiquement exposés ainsi que sur la portée éthique et sociale de ce type de solutions techniques.

Références

Androutsopoulos, I., G. Lampouras, et D. Galanis (2013). Generating natural language descriptions from OWL ontologies : the naturalowl system. *J. Artif. Intell. Res. (JAIR) 48*, 671–715.

Banik, E., C. Gardent, et E. Kow (2013). The KBGen Challenge. In *the 14th European Workshop on Natural Language Generation (ENLG)*, Sofia, Bulgaria, pp. 94–97.

Bontcheva, K. (2005). Generating tailored textual summaries from ontologies. In *The Semantic Web : Research and Applications, Second European Semantic Web Conference, ESWC 2005, Heraklion, Crete, Greece, May 29 - June 1, 2005, Proceedings*, pp. 531–545.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Commun. ACM 13*(6), 377–387.

data.gouv.fr, Plateforme ouverte des données publiques françaises (2016). Sélection thématique : élections régionales 2015. `https://www.data.gouv.fr/fr/datasets/selection-thematique-elections-regionales-2015/`. Accessed : 2016-10-14.

Demir, S., S. Carberry, et K. F. McCoy (2012). Summarizing information graphics textually. *Computational Linguistics 38*(3), 527–574.

Dumontier, M., L. Ferres, et N. Villanueva-Rosales (2010). Modeling and querying graphical representations of statistical data. *J. Web Sem. 8*(2-3), 241–254.

Labbé, C. et F. Portet (2012). Towards an abstractive opinion summarisation of multiple reviews in the tourism domain. In *The First International Workshop on Sentiment Discovery from Affective Data (SDAD 2012)*, pp. 87–94.

Labbé, C., C. Roncancio, et D. Bras (2015). A Personal Storytelling about Your Favorite Data. In *15th European Workshop on Natural Language Generation (ENLG 2015)*, Brighton, United Kingdom.

McKeown, K. R. (1985). *Text Generation : Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press.

Portet, F., E. Reiter, A. Gatt, J. Hunter, S. Sripada, Y. Freer, et C. Sykes (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence 173*(7-8), 789–816.

Reiter, E. et R. Dale (2000). *Building Natural Language Generation Systems*. New York, NY, USA : Cambridge University Press.

Syllabs (2016). Nos robots rédacteurs collaborent avec le monde. http ://blog.syllabs.com/le-monde-elections-departementales-syllabs-robotjournalisme/. Accessed : 2016-10-14.

Vaudry, P.-L. et G. Lapalme (2013). Adapting simplenlg for bilingual english-french realisation. In *Proceedings of the 14th European Workshop on Natural Language Generation*, Sofia, Bulgaria, pp. 183–187. Association for Computational Linguistics.

Summary

Natural language generation is used to describe, in natural language, the data answering a user query : weather forecast, elections or sport results... In most cases, the generated text do not show the singularity nor the normality of the selected data compared to the whole. Most of the time, the knowledge needed to express these particular kind of information is available. It is thus possible to detect particular exceptional or banal features of a selection and to generate automatically a text that exposes these interresting facts. This paper presents a tool aiming at specifying generators of texts containing comparisons of the selected data to the whole set. Singularity and banality exprimable using the tool are showed thanks to an example (election results). Reification of the needed knowledge (models, language ressources,...) aims for genericity of the approche and easy reuse for other domains (weather forecast, sports,...). The prototype `Summy` was built to validate the approach and to demonstrate how particularity or normality of a subset of data compared to the whole can be automatically expressed.

Analyse des Media Français: Quand l'économie rencontre la fouille de donnée

Marie-Luce Viaud, Nicolas Hervé*, Julia Cagé**

*INA, 4 avenue de l'europe, 94366 Bry Sur Marne
mlviaud@ina.fr, nherve@ina.fr
**Sciences Po, rue des Saints Pères, 75005 Paris
julia.cage@sciencespo.fr

Résumé. Après la mutation numérique, quel visage présentent les media français ? Analyse quantitative de la propagation de l'actualité, sur l'année 2013 et dans les principaux media Français en ligne (presse, pure player, radio, TV et AFP), cette recherche associe des outils de fouille de données et d'économétrie. Les évènements médiatiques les plus importants se propagent en moins de 5 minutes, avec un taux de reprise élevé. Nous montrons que l'audience ne bénéficie que dans une tès faible mesure aux "news breakers". Face à ces résultats, nous élaborons quelques pistes de reflexion sur les enjeux et l'évolution des modèles économiques des média.

1 Introduction

Une étude qui rassemble "Media Studies" et "Big Data", deux termes porteurs de cette décénie où le volume des "traces" permet d'espérer le graal d'une "approche quantitative", même dans des domaines jusque là, le plus souvent, réservés au qualitatif. 5 années de travail - certes pas à temps plein mais néanmoins-, pour valider les données, développer les algorithmes, dépouiller les résultats, contourner les biais, développer des hypothèses et les confronter aux données et au contexte réels et enfin développer les interpretations et les enjeux liés au contexte actuel.

1.1 Propagation de l'information à l'heure de l'internet

Schumpeter et Bourdieu l'ont constaté, et vous aussi sans doute : la propagation de l'information est circulaire dans bien des cas. Reste qu'à l'heure d'internet, le cercle semble se resserrer dangereusement. Dangereusement pour qui ? Pour les media qui perdent des lecteurs ? Pour les journalistes qui perdent leur crédibilité ? pour les citoyens lecteurs qui perdent leur confiance ? Pour la société démocratique qui perd une diversité d'opinion et de point de vue qui la constitue et lui donne sa vitalité ?

1.2 Un constat à la base de cette étude : les media sont en difficulté

1993 voit la naissance du web. Plus qu'une mutation technologique, internet et le web impactent à la fois la production, la diffusion et la consommation des media. De 1995-2000 les media investissent dans le online : une diffusion immédiate pour un relais de l'information instantané. Les media gagnent en réactivité et en visibilité. Mais les modèles économiques sont mis à mal : l'information en ligne est difficile à monétiser. Avec la gratuité de consommation en légende, les nouveaux usagers de l'internet ne sont pas décidés à payer l'information, qu'elle soit journalistique on non. La part des revenus publicitaires allouée aux media traditionnels n'a cessé de diminuer depuis 15 ans pour descendre en 2014, à moins de 29% du revenu total (Charon, 2015). Quant aux revenus publicitaires numériques, absorbés par les infomédiaires , ils atteignent péniblement actuellement 7% de ce même revenu total (Mitchell et Holcomb, 2016). Cette morosité économique n'est pas sans conséquences. Entre 2000 et 2010, plus de 140 media font faillite aux Etats-Unis. Depuis le début du siècle et dans tous les pays sauf le japon, la taille des rédactions s'érode de manière continue, pour atteindre dans certains cas jusqu'à 30% de réduction des effectifs. En outre, depuis plus de 30 ans, la diffusion des quotidiens populaires et certains régionaux régresse fortement.

Utilisant les données collectées par le projet ANR OTMedia, cette étude se propose d'observer en détail la production, la diffusion et la consommation des media en tentant de répondre aux questions suivantes :

— Peut on mesurer un taux de reprise de l'information à partir des données collectées ?
— Le taux de reprise d'information est-il corrélé à la taille de la rédaction d'un journal ?
— La reprise d'information fait -elle l'objet de référencement de la part des *emprunteurs* ?
— Le media qui publie l'information originale en beneficie-t'il directement en terme d'audience ?
— En bénéficie-t'il indirectement ?
— Quels sont les acteurs et les enjeux du débat ?
— quelles pistes de reflexion peuvent se dégager de cette étude quantitative pour assurer la production d'une information de qualité ?

2 Les données

Le projet OTMedia (ANR-Content 2010-2013) nous a permis de collecter l'ensemble des media français sur l'année 2013. Notre étude se focalise plus spécifiquement sur le contenu produit en ligne par les media d'information générale et politique. Il est composé de 86 media d'actualité : l'Agence France Presse (AFP), 59 journaux (35 titres de presse quotidienne régionale, 7 titres de presse quotidienne nationale, 3 gratuits, 12 hebdomadaires nationaux et 2 mensuels), 9 télévisions, 7 radios et 10 pure players (produisant uniquement de l'information en ligne). Pour chacun de ces médias, nous avons collecté les articles publiés en ligne en 2013 (que ces articles soient disponibles à tous ou derrière un mur payant) en utilisant leurs flux RSS principaux et/ou leurs Sitemaps. Il a été complété à posteriori par de nouvelles captures pour les médias qui n'étaient pas couverts dans le cadre d'OTMedia, notamment Ouest-France. Notre corpus comprend 2 548 634 documents pour l'année 2013, soit en moyenne environ 7 000 documents par jour. La taille moyenne de ces documents est de 1 865 caractères. 73% de ces documents proviennent des sites internet de la presse papier (57% des journaux locaux et

près de 16% des journaux nationaux), 6,6% de sites internet de chaînes de télévision, 4,7% de sites internet de stations de radio, 12,5% de l'AFP et 3,2% des pure players En sus, nous avons collecté les données économiques et d'audience relatives à chaque media de la base : en particulier la taille de la rédaction et son audience journalière online pour l'année 2013 (Données OJD).

3 La boite à outils

Notre objectif est de retracer le parcourt de l'information au travers des différents media. il s'agit donc de mettre en oeuvre la collecte de l'information, l'extraction des contenus des articles, et de retracer les reprises et les référencements. Néanmoins, les reprises et référencements sont pertinents dans le cas d'un même fait propagé par plusieurs media.

3.1 Détecteur d'événements

L'événement mediatique est un concept très étudié et discuté par plusieurs communautés de chercheurs. Les sciences sociales approchent l'événement mediatique par toutes ses facettes : interrogation sur les faits à la base de l'évènement, sur la construction par les media, la reception par le public et les conséquences sur la société (Neveu et Quéré, 1996). Cette définition sort du cadre de ce qui peut être mesurable dans notre contexte. Aussi avons-nous défini la notion d'évènement mediatique comme un agrégat "sufisamment important" de documents relatant un même fait. Bien qu'extrêmement réductrice, cette définition est déjà relativement floue, car il est difficile de délimiter même manuellement les "bornes" d'un événement. Lors de la réalisation de vérités de terrain, nous avons constaté que deux annotateurs ne produisaient pas exactement les mêmes événements, certains événements pouvant être scindés en plusieurs événements. Par exemple, l'affaire Cahuzac est constituée d'un certain nombre de rebondissements, les jeux olympiques sont constitués d'autant d'événements que de médailles... La détection d'évènement se base sur les travaux en "topic detection" (Allan et al., 2006) et comporte 6 phases :

- pré-traitements : suppression des mots vides et stemmatisation.
- les vecteurs TF-IDF sont calculés pour chaque document. Les mots apparaissant dans le titre disposent d'un facteur mumtiplicatif de 5. Ce facteur été déterminé comme optimal après l'utilisation du moteur de recherche OTMedia pour des études pendant près de deux années.
- une similarité cosinus est calculée entre chaque vecteur-document
- un algorithme d'aglomération itérative construit des clusters de vecteurs documents, un seuil pour l'aglomération a été fixé à partir d'une vérité de terrain manuelle.
- Nous considérons qu'un événement est finalisé s'il ne reçoit plus de nouvel article pendant une certaine période de temps. Nous utilisons une fenêtre glissante de 24 heures. 6. Pour qu'un cluster soit labelisé "événement", il faut qu'il contienne au moins 10 documents issus de deux média différents.

Le détecteur d'évènement a été testé sur le corpus test de la communauté (TDT Pilot Study Corpus). Deux types d'algorithmes sont evalués dans cette campagne : retrospectifs et online. la version retrospective considère l'ensemble des documents de l'année pour effectuer le clustering alors que l'algorithme online prend les articles au fur et à mesure de leur apparition pour

les attribuer à un cluster. Pour favoriser l'évolution de notre plateforme, nous avons choisi de développer un algorithme online. L'évaluation de l'algorithme proposée est basée sur 15000 articles alors que notre système considère plusieurs millions de documents. Deux mesures sont utilisées pour l'évaluation officielle : les moyenne F1 macro et micro. Les résultats sont les suivants :

Run	micro-avg-F1	macro-avg-F1
Retrospective CMU incremental	0.64	0.77
Online CMU decay - win2500	0.40	0.39
Notre implementation	0.55	0.69

Notre implementation obtient de meilleurs résultats pour une implementation online que l'état de l'art. est n'est que de 10% inférieure aux resultats relatifs à une implémentation retrospective. Nous avons réalisé un deuxième test avec les données de l'European Media Monitoring sur l'année 2013, et nous retrouvons aussi 90% des évènements.

3.2 Détecteur de référencement

Le détecteur de référencement cumule une recherche de formules de référencement classiques pour les media comme "@ AFP-2013" ou "source Reuters" et du détecteur syntaxique Unitex (Gross, 1997), auquel nous avons fourni des patrons syntaxiques spécifiques. Nous avons effectué des vérités de terrain afin d'évaluer la qualité du détecteur. 95% des détections de référencement effectuées sont correctes (précision). Néanmoins, seulement 80% des citations de media non repertoriées comme des référencements, n'en sont effectivement pas (rappel).

3.3 Détecteur de copies

Les algorithmes qui détectent les portions de textes identiques entre deux documents existent depuis très longtemps en informatique. Ils sont un composant essentiel de nombreux autres traitements. Nous parlons bien ici de copié-collé et non pas de plagiat, notion plus générale qui implique une reformulation potentielle du texte d'origine. Nous avons développé une implémentation nous permettant de trouver rapidement, pour un article donné, l'ensemble des articles d'un corpus avec lesquels il partage du contenu. Afin d'accélérer les traitements, nous utilisons une technique de hachage (hashing) qui permet de représenter une donnée à l'aide d'une empreinte particulière. C'est cette empreinte qui sert à identifier rapidement les contenus similaires. Dans notre cas, nous avons « haché » les suites de cinq mots consécutifs (représentation classique en TAL, dénommée n-gram, et donc 5-grams dans notre implémentation). Nous pouvons ainsi repérer les contenus identiques à partir du moment où ils sont constitués au minimum de cinq mots. Pour tout couple de textes présentant du contenu identique, nous agglomérons toutes ces suites de cinq mots pour former les blocs de texte identique. Pour considérer que deux textes présentent de la copie, nous imposons un seuil minimal de 100 caractères sur la taille d'au moins l'un de ces blocs1.

En appliquant cet algorithme au corpus formé par l'ensemble des articles d'un événement médiatique, nous pouvons mesurer plusieurs choses. En prenant les articles dans leur ordre chronologique, nous pouvons identifier pour chacun d'entre eux la proportion de contenu qui

est originale et la part qui est copiée. Pour cette part copiée, nous distinguons ce qui est copié au sein du média (réutilisation de sa propre production) de ce qui est emprunté à un autre média. De plus, nous sommes également capables de dire dans quel article un bloc de texte est apparu pour la première fois et ainsi d'en attribuer la paternité au bon média.

On peut bien évidemment discuter de la « validité » de ce seuil de 100 caractères, que l'on pourrait vouloir un peu plus long ou bien un peu plus court. Le raccourcir, c'est prendre le risque de qualifier improprement de copié-collé des tournures qui sont simplement usuelles dans l'écriture journalistique. Le rallonger au contraire, c'est prendre le risque de ne pas considérer comme du copié-collé du contenu qui est effectivement du copié-collé. Ce qui nous intéresse principalement ici ce n'est pas tant la quantité de contenu copié-collé que la quantité de contenu original produit par le média, et que modifier faiblement ce seuil ne modifierait pas quantitativement les résultats que nous obtenons.

3.4 Régression

A cela s'ajoute les outils des économistes : les régressions simples et multiples, linéaires ou polynomiales. Il s'agit d'approximer la variable "expliquée" y^* à l'aide des variables explicatives $x_1, x_2, ...x_n$, selon une fonction $y^* = f(x_1, x_2, ...x_n)$. D'un point de vue interprétatif, si l'erreur $|y_{reel} - y^*|$ est assez faible ; alors l'approximation par f est déclarée valide. La corrélation peut être interprétée comme un effet causal à deux conditions : si l'effet de corrélation est bien identifié de x sur y et non l'inverse, et si on n'omet pas de variables explicatives.

Les regressions multiples ont été utilisés pour étudier la corrélation entre la taille de la rédaction (nombre de journalistes), la production d'information (totale, événementielle et originale) et la variation de l'audience. Pour déterminer la corrélation entre la taille de la rédaction et le nombre de journalistes, on introduit : les variables explicatives de f : (1) le jour de la semaine (2) le jour de l'année. Pour déterminer la corrélation de la production d'information originale sur l'audience des sites internet des médias, nous avons régressé la part d'audience quotidienne de l'ensemble des médias de notre corpus sur un certain nombre d'« effets fixes jour » mais également d'« effets fixes média » : (1) le contenu total non classé dans des événements ; (2) le contenu total classé dans des événements ; (3) le contenu original ; et (4) le nombre d'événements pour lesquels le média est le « news breaker ».

4 Confrontation des résultats, des données, des concepts et des biais

Chaque résultat a fait l'objet de discussions et de retour vers les données pour les valider ou trouver d'autres angles d'approche pour les mettre en default. C'est un travail itératif, chaque biais est étudié pour vérifier qu'il n'amplifie pas les résultats dans le sens de l'interprétation.

4.1 Des données horaires surprenantes

Le marquage horaire des media par eux-mêmes semble une source dont la fiabilité dépend ...du media. En effet, entre les changements d'heures (été/hiver) répercutés avec du retard ou pas du tout, les erreurs ponctuelles flagrantes (la mort de Steeve Job annoncée un jour à

l'avance !), les différence de marquage entre le flux RSS et l'article lui même, des pratiques éditoriales quelque peu relachées ou l'oubli de référencement du media original, suivre la propagation de l'information n'est pas aussi simple qu'il le semblerait au premier abord ! Devant des résultats pour le moins surprenants, nous avons dû développer des méthodologies spécifiques :

— Détecter les media qui ne semblaient pas connaître le changement d'heure et les redater
— detecter les media qui référençaient un autre media en en copiant une partie tout en publiant avant lui.
— étudier la répartition des documents dans les 10 premières minutes et traquer les media systématiquement en avance.

Dans cette étude, il est fondamental d'avoir une bonne précision sur le "news breaker", et les outils tout automatiques ont ici montrés leurs limites. Aussi, pour finir, avons-nous réalisé une interface permettant de visualiser les deux documents et leur contenu partagé, afin de déterminer manuellement lequel des deux documents est l'original. Si le choix est impossible, le couple de document est labelisé "indéterminé". Cette interface tire aléatoirement un couple de documents ambigüs au niveau date et présentant un taux de copies élevé (>50%). Elle permet d'étudier les fonctions de "correction horaire" à apporter aux différents médias. Les couples "indéfinis" seront ordonnés dans le temps aléatoirement afin de répartir le biais sur l'intégralité de la distribution.

4.2 une étude sans bruit ?

La segmentation du web n'est pas parfaite. Sur certains sites, la captation des contenus inclut des parties d'habillage ou d'éléments structurels de certains sites (Abonnez vous pour ..., LES TITRES DU JOUR..., @AFP : l'usage de ce contenu...) constituées de blocs de textes spécifiques à chaque media. Certes, nous travaillons déjà sur une segmentation plus propre, en détectant les parties communes récurrentes intra-media et en les validant comme bruit. Mais l'important est d'estimer en quoi ce biais influe sur l'étude. Ces contenus d'habillage vont avoir tendance, de manière générale, à augmenter dans une faible mesure le volume d'information. Les taux de copie intra-media et d'originalité inter-media seront impactés dans une plus forte mesure. Néanmoins, les valeurs importantes pour notre propos sont le volume de production et le taux de copie inter-media : le danger ici serait de sous-estimer le volume de production et de surestimer le taux de copie inter-media. Or ce biais a l'effet contraire : nous sur-estimons faiblement le volume produit et sous-estimons le taux de copie inter-media.

4.3 "news breaker", un phénomène qui recouvre plusieurs réalités

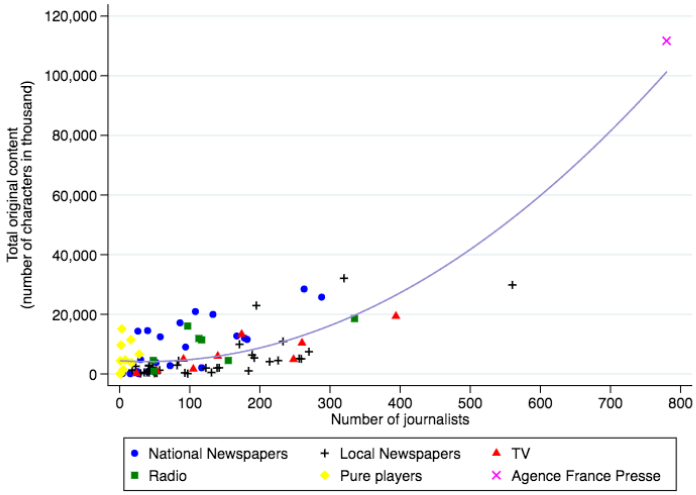
La notion de news breaker, premier media sur un événement, est importante pour notre propos. Néanmoins, elle recouvre des réalités assez diverses. Si l'évènement est le résultat d'une investigation, comme les panamas papers ou l'affaire Cahuzac par exemple, alors le "news breaker" est réellement le media qui "sort " l'affaire. Mais si l'article est issus d'une conférence de presse à laquelle assiste plusieurs media, il est difficile de savoir qui copie qui, et le rôle même de news breaker est plus flou, puisque tous les journaux reçoivent la même information en même temps par un acteur tierce. Ces constatations nous ont conduit à affiner notre approche en classifiant les événements en 8 catégories dont 2 concernant les événements exclusifs à un média et 6 pour les événements non exclusifs (résultats sportifs, politiques,

évènement culturels ...) afin de préciser ce concept de news breaker et affiner les interprétations qui en découlent.

5 Les résultats

5.1 Production et taille des rédactions : une corrélation significative

Les corrélations entre le nombre de journalistes et la production d'information, d'événements et d'information originale sont positives et statistiquement significatives (cf figure1). Les valeurs de corrélation sont plus élevées pour la production d'information classées dans des événements et pour l'information originale que pour la production d'information contextuelle ou les articles d'actualité locale non classés dans des événements. En moyenne annuelle, une augmentation de 1% du nombre de journalistes travaillant dans une rédaction augmente le contenu non classé dans des événements de 0,78%, le contenu classé de 1,29% et le contenu original de 1,20%. De plus, une augmentation du nombre de journalistes de 1% augmente le nombre de « breaking news » d'un peu plus de 1%. Si l'on réfléchit en termes de nombre de



Notes: The Figure shows the correlation between the total original content produced by each media outlet in 2013 and the number of journalists. Source: Cagé et al. (2016).

FIG. 1 – *Corrélation entre la production originale et le nombre de journalistes dans la rédaction*

journalistes, nos résultats montrent qu'un journaliste supplémentaire dans une rédaction augmentera en moyenne de 28 articles entièrement originaux par an la production d'information classée dans des événements. La différence entre le nombre de journalistes du Monde et le nombre de journalistes de Libération possédant la carte de presse en 2013 – soit 110 cartes de presse – permet à elle-seule d'expliquer près de la moitié de la différence quantitative de

production d'information originale entre ces deux médias (environ 12 millions de caractères originaux pour Libération en 2013 contre près de 26 millions pour Le Monde). Cette augmentation de la production provient-elle du fait que les médias aux rédactions les plus grandes couvrent plus d'événements, et/ou du fait qu'ils publient plus d'articles (ou des articles plus longs) à l'intérieur des événements qu'ils couvrent ? Pour répondre à cette question, nous avons étudié la production d'information par les médias à l'intérieur de chaque événement . Nous trouvons que, si la production d'information à l'intérieur des événements augmente avec le nombre de journalistes, la plus grande partie de l'effet provient du plus grand nombre d'événements couverts par les médias ayant plus de journalistes. Autrement dit, les plus grandes rédactions couvrent plus d'événements ; à l'intérieur de chaque événement, elles produisent également plus d'information, et en particulier plus d'information originale, mais cet effet est quantitativement moins important.

5.2 Pas d'influence du support d'origine

Les médias ne diffèrent pas « en ligne » en fonction de leurs supports « hors ligne ». Cette réalité, qu'il est important de rappeler, peut s'expliquer en partie par le fait que sur Internet, les médias ne diffèrent que relativement peu. Même le direct vidéo n'est plus restreint aux sites des télévisions et des radios lors de périodes médiatiques extrêmement actives comme des attentats ou des élections, comme en atteste le "live" du Monde. Les médias nationaux et médias locaux présentent, quant à eux, des différences notables. À taille de rédaction comparable, ils génèrent un volume d'information relativement semblable. Néanmoins, le nombre d'évènements et la proportion d'articles classés dans des événements est beaucoup plus faible pour les media locaux. Ce constat s'explique assez aisément par notre définition même de la notion d'évènement puisqu'il faut qu'au moins deux médias et une couverture de 10 articles sur un même fait pour produire un évènement. Et le plus souvent, les journaux régionaux sont confrontés à une concurrence limitée sur leur territoire, d'où la difficulté pour un fait local d'atteindre la cathégorie "événement".

5.3 AFP : un rôle spécifique

De par son statut, son rôle, sa taille, l'étendue de sa couverture médiatique ou son volume de production, l'AFP se distingue clairement des autres acteurs. Le modèle économique de l'afp est un modèle BtoB d'agence de presse : ses dépêches ne sont pas directement accessibles au grand public. Sur l'année 2013, L'AFP représente 12% de la production d'actualités en ligne. L'AFP est ainsi quantitativement l'un des principaux producteurs d'information en France, seulement devancée par Ouest France et ses très nombreuses éditions locales. Avec une rédaction de plus de 750 journalistes, l'AFP est "news breaker" sur la moitié des événements et couvre 93% des événements. Mais Est-ce à dire que sa production se focalise sur les évènements ? Sur notre corpus, seulement la moitié des dépêches produites sont classées dans des évènements, montrant que malgré une couverture très forte de l'actualité "vive", l'AFP n'est pas en reste sur les sujets moins "chauds".

5.4 Une vitesse de propagation extrêmement élevée

La vitesse de propagation d'un événement en ligne est extrêmement rapide. Il faut en moyenne moins de trois heures (175 minutes) pour qu'un deuxième média couvre un événement déjà couvert par un premier média. Trois heures en moyenne cela peut paraître long, mais cette statistique recouvre une forte disparité. En effet, un certain nombre d'événements se propagent beaucoup plus rapidement. La moitié des événements se propagent en moins de 25 minutes ; un quart des événements se propagent en seulement... 230 secondes, et 10% en seulement 4 secondes, ce dernier chiffre reflétant la publication automatique sur le site d'un certain nombre de médias de contenus déjà mis en forme par l'AFP.

5.5 La reprise, un usage ... très répandu

Nous definisson le taux moyen d'originalité d'un article comme la part du contenu original de l'article rapporté à son contenu total. le taux d'originalité mesuré est de 36%. En conséquence, 64% de ce qui est publié en ligne est du copié-collé sans prendre en compte les reformulations. Comme le montre la figure (cf figure 2) la distribution du taux d'originalité est bimodale, près de 19% des articles classés dans des événements ne présentent aucune originalité, et 21% d'entre eux sont entièrement originaux . Plus de la moitié des articles (56%) comprennent moins de 20% de contenu original.

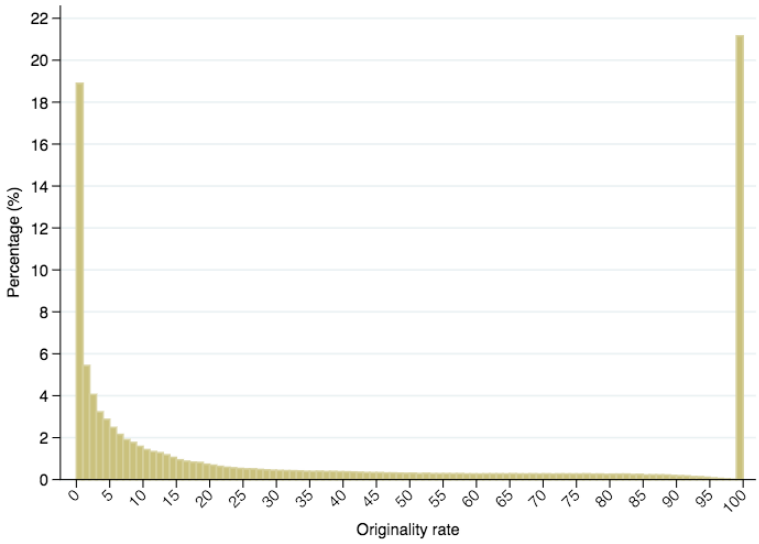


FIG. 2 – *Taux d'originalité sur tous les documents classés dans des événements*

Sur les 850 000 documents classés dans des événements, plus de 620 000 (73%) présentent de la copie externe. Le taux moyen de copie externe est de 56% ; 74,8% conditionnellement au fait d'avoir recours à la copie. Le reprise sous forme de copier-coller peut donc être qualifiée

de conséquente du point de vue du média qui copie, mais qu'en est-il du media copié ? nous mesurons que 72% du contenu du média repris est copié : en d'autres termes, les médias d'information reprennent directement près des trois quarts des articles originaux. Il est important de souligner à nouveau ici que notre analyse repose uniquement sur l'étude du contenu produit par les médias d'information. Ne sont pris en compte ni les documents publiés sur les réseaux sociaux ni ceux publiés par les agrégateurs. Pour relativiser ces chiffres, rappelons néanmoins que les événements représentent moins de la moitié de la production. Même si cela revient sans doute en partie à la surestimer , on peut penser que l'originalité de ces articles de long court, qui ne s'inscrivent pas dans les événements, est beaucoup plus proche de 100%. Ce qui conduit à relativiser l'importance du copié-collé sur l'intégralité de la production.

5.6 Originalité et audience des sites de media

Les résultats que nous présentons ici sont surprenants. Le seul déterminant causal statistiquement significatif de la part d'audience des médias est la production de contenu original. Le fait d'être « news breaker » n'affecte pas l'audience en ligne des médias. Cependant, s'il est statistiquement significatif, l'effet de la production originale d'information est néanmoins quantitativement faible : en moyenne quotidienne, une augmentation de 1% de la production de contenu original n'augmente que de 0,016% le nombre de visiteurs uniques d'un site. Autrement dit, les utilisateurs ne « récompensent » presque pas les producteurs d'information originale (ni ne punissent les « copieurs »). L'information en ligne étant substituable – tous les sites reprenant simultanément les mêmes informations souvent avec les mêmes mots – celui qui la produit ne parvient donc plus à la monétiser, tout du moins en termes d'audience.

5.7 Référencement

Tout d'abord, 42,5% des articles classés dans les évènements contiennent des referencements à d'autres media, contre seulement 14% pour les articles non classés. Ces chiffres semblent cohérents avec le taux de copie conséquent des documents classés, et le fait que les documents non classés auraient des contenus plus "indépendants". Nos algorithmes ont détecté 461 336 citations en 2013, dont 48% sont des références à l'AFP. La surreprésentation de l'AFP s'explique en partie par le nombre élevé d'articles écrits « avec AFP » ou @AFP2013. 11% des citations se rapportent à des médias étrangers comme Associated Press, Reuters, le Guardian ou encore le New York Times. En outre, certains media cités ne font tout simplement pas partie de notre base de donnée, comme tout particulièrement le Canard Enchaîné, qui ne dispose pas de site web. En moyenne, chaque media inclu dans notre base de données est cité un peu plus de 4 500 fois en 2013 , 1 900 fois si l'on ne considère par l'AFP. Les radios et les télévisions arrivent en tête des médias les plus cités avec plus de 12 700 citations pour RTL, près de 12 000 pour Europe 1 et 9 000 pour BFM TV. Une première analyse sommaire de ces référencements semble montrer que la plus large partie provient des interviews réalisés par les media audiovisuels et radiophoniques, avec notamment des reprises importantes des matinales radio. Viennent ensuite, avec des scores très proches, les reprises des journaux Le Monde, Le Parisien, Le Point ou encore Le Figaro. Enfin, les pure players et les journaux locaux sont parmi les moins cités. Le Dauphiné Libéré, La Provence, Nice Matin, Ouest France, Le Progrès, L'Union, Le Courrier Picard et Sud Ouest, par ordre croissant de citations reçues, ont été cités plus de 1 000 fois en 2013.

6 Conclusion

6.1 Manque d'incitation pour une information originale ?

Les coûts fixes des journaux sont importants, principalement les couts redactionels, et ne varient pas avec leur audience. C'est d'autant plus vrai à l'ère numérique, où impression, et routage n'ont plus lieu d'être. Historiquement, Ces coûts fixes élevés étaient assumés par les media, car ils espéraient, en contrepartie, augmenter leurs ventes et leurs profits avec leurs scoops et exclusivités. Publier une information en exclusivité pour un journal papier lui assurait, au moins durant une journée, une augmentation de ses ventes et donc de ses revenus, car cette information n'était pas disponible dans les journaux papier concurrents. Comme nous l'avons vu précedemment, l'extrême rapidité de la diffusion des nouvelles du à l'internet confisque ce profit direct du news breaker . Se pose alors la question de l'incitation à produire de l'information originale, de fait coûteuse. Certes la réputation d'un média est issue, en partie, du bénéfice cumulé des informations originales produites par ce media. Mais encore faut-il que les crédits soient bien attribués aux news breakers. Or, nos résultats montrent que les référencements des média "lanceurs" hors AFP restent assez faibles (21% des contenus regroupés sous forme d'événements). Cet aspect sera étudié plus en détail lors de travaux ultérieurs.

6.2 Crowd sourcing, Copyright, droits voisin, syndication et murs payants : innovations et tentatives de monétisation directe

Quelles possibilités pour les media de monétiser directement leur valeur ajoutée ? Dans ce contexte difficile les media répondent par un dynamisme exploratoire assez extraordinaire : de nouveaux media pure players apparaissent depuis deux ans, proposant des approches innovantes en terme de contenu : "8ième étage" avec son slogan « différencier l'information de l'actualité », "Les Jours" avec ces articles reportages, "Contexte" specialisé dans l'information européenne, "Ijsberg", "L'imprévu", "Brief.me" qui propose une information personnalisée. Sans compter le "1", "Society" ou encore "Soixante Quinze" en publications "papier". Certains misent sur le crowd sourcing et la "responsabilisation" des lecteurs pour obtenir une information de qualité. Néanmoins, d'autres solutions peuvent être discutées. Si la monétisation indirecte sous forme de publicité se restreint comme peau de chagrin (Mitchell et Holcomb, 2016), l'emprunt systématique de contenu peut amener à questionner le bien fondé d'une réapparition du copyright ou droits voisins pour le domaine de l'information. Par ailleurs devant ces problèmes de moyens de plus en plus pregnants, la grande tendance des acteurs médiatiques est actuellement mettre en place des murs payants. Le NewYork Time, Le Monde, le Time, le Wall Street journal, le Figaro, les Echos ont dejà franchi le pas, sans voir pour autant leur lectorat s'éffrondrer. Enfin la syndication de contenu permet aux journaux de référence de générer des revenus non négligeables (5 Millions d'euros l'an dernier pour le financial times selon un article de la trubune). La syndication de diffusion, si elle est pour le moment restreinte aux nouveaux media, permettrait, tout en contournant ponctuellement la voracité des infomédiaires, d'envisager de nouveaux modes de consommation. Le lecteur aurait accès à l'article à la demande dans une offre multiple, soit un choix informationnel adapté à un usage instantané, borné par le revenu ou le temps disponible.

6.3 Quid de la complétude et de la diversité de l'information ?

Car il nous faut également questionner notre relation de citoyens consommateurs – et pourtant si souvent mauvais payeurs – à l'information. Quelle information souhaitons-nous collectivement et à quel prix ? Secteur d'activité à part dans le tissu économique, média et information tiennent une place essentielle pour assurer la vitalité de nos démocraties. On est prompt à souligner leurs manquements, encore faudrait-il rappeler leur utilité. On ne peut éluder cette question. L'objectif à poursuivre est bien d'assurer une information la plus complète possible des citoyens, de garantir l'expression de la diversité des points de vue et de refuser un modèle qui diffuserait plus en produisant moins.

Références

Allan, J., . Harding, D. Fisher, A. Boliver, S. Guzman-Lara, et P. Amstutz (2006). Taking topic detection from evaluation to practice. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05))*.

Charon, J.-M. (2015). Presse et numérique - l'invention d'un nouvel écosystème. *http ://www.culturecommunication.gouv.fr/Ressources/Rapports/Rapport-Charon-Presse-et-numerique-L-invention-d-un-nouvel-ecosysteme*.

Gross, M. (1997). The construction of local grammars. *in E.Roche et Y.Schabes (eds.), Finite-State Language Processing, Cambridge, Mass./London, The MIT Press*, 329–352.

Mitchell, A. et J. Holcomb (2016). State of the news media 2016. *http ://www.journalism.org/2016/06/15/state-of-the-news-media-2016/*.

Neveu, E. et L. Quéré (1996). Le temps de l'événement. *Revue Réseaux, numéro 75, CNET*.

Summary

This work presents an analysis of news propagation on online french media. The data base is a dataset including all online content produced by the universe of news media (newspaper, television, radio, pure online media, and a news agency) in France during year 2013. We develop algorithms for events, copy and past and references detection to trace the timeline of each story. The most important events propagate in less than 5 minutes, with an important rate of copy and paste. Moreover, we show that breaking news only slightly increases the number of online viewers. From these results, we draw up some propositions about the evolution of media economic models.

Étude des influences réciproques entre médias sociaux et médias traditionnels

Béatrice Mazoyer*, Nicolas Turenne**, Marie-Luce Viaud*

*INA, 18 Avenue des frères Lumière, 94366 Bry-sur-Marne, France
bmazoyer@ina.fr, mlviaud@ina.fr
**Université Paris-Est, LISIS, INRA, 77454 Marne-La-Vallée, France
nturenne@u-pem.fr

Résumé. Cet article s'intègre dans un travail de recherche consacré à l'étude des médias traditionnels et des réseaux sociaux et de leurs influences mutuelles. Notre étude vise à mettre en place un outil de collecte en continu des tweets liés aux événements relayés par les médias traditionnels. L'objectif est d'obtenir pour chaque événement médiatique un corpus à la fois exhaustif et précis (minimisation du bruit) de tweets qui fassent référence à cet événement. Notre méthodologie se fonde sur un processus itératif : les tweets collectés dans un premier temps sont analysés pour affiner les collectes suivantes. Notre outil obtient de bons résultats en ce qui concerne la pertinence des tweets collectés vis-à-vis des événements, mais l'exhaustivité semble encore à perfectionner, ce pourquoi nous proposons des pistes d'amélioration.

1 Introduction

Les réseaux sociaux jouent de façon croissante le rôle d'intermédiaires entre les médias traditionnels (presse, radio, télévision) et leur audience. Les médias français s'adaptent à cette évolution : ainsi, de nombreux médias, parmi lesquels Le Parisien, Libération, FranceTVInfo ou L'Équipe, ont adopté en 2016 le format Facebook Instant Article [1] (qui permet une lecture facilitée sur mobile, sans quitter l'application Facebook) quitte à perdre une partie de leur contrôle sur leur format de distribution.

À cela s'associe une évolution des contenus publiés : les médias traditionnels s'inspirent des modes de communication des médias sociaux et relaient les informations les plus partagées sur ceux-ci. Cependant, il est difficile de quantifier et d'analyser de façon précise les influences réciproques entre ces deux sphères, notamment car on ne dispose pas d'un jeu de données associant événements médiatiques et réactions sur les réseaux sociaux sur une longue durée. C'est l'enjeu du travail que nous présentons ici : la construction d'un outil associant automatiquement à des événements d'actualité les tweets qui y font référence. Cet article présente des travaux en cours : nous travaillons actuellement sur la formalisation de l'évaluation des résultats, ce qui représente une tâche complexe car s'il est possible d'annnoter manuellement les tweets récoltés par l'outil, il est plus difficile d'estimer le volume de tweets non détectés.

1. https ://developers.facebook.com/docs/instant-articles

Le choix de Twitter comme média social étudié doit être justifié, étant donné la prédominance de Facebook en termes de nombre d'utilisateurs [2]. Pour notre étude, Twitter a l'intérêt d'être un réseau social où les interventions sont publiques, à l'inverse de Facebook, où la plupart des utilisateurs retreignent la visibilité de leurs publications à un cercle privé. C'est probablement la raison pour laquelle beaucoup de journalistes utilisent de façon privilégiée Twitter comme outil de travail [3]. Dans le cadre d'un travail sur les influences des médias, l'étude de Twitter nous a donc parue fondamentale.

2 Travaux antérieurs

D'autres travaux de recherche se sont consacrés à l'étude des liens entre tweets et événements médiatiques. Beaucoup travaillent cependant sur des corpus de tweets fixés, constitués au préalable, qu'ils cherchent à répartir en « sujets » ou « événements » décrits par les médias, soit de façon manuelle (Rieder et Smyrnaios, 2012), soit en utilisant des modèles thématiques de type LDA (Zhao et al., 2011; Hu et al., 2012; Hua et al., 2016). À l'inverse, notre démarche consiste à collecter en continu de nouveaux tweets associés à des événements d'actualité. Il s'agit d'une approche dynamique du traitement de l'information.

Des méthodes de type « First Story Detection » sont plus adaptées à des documents collectés en continu : elles consistent à attribuer chaque nouveau document d'un flux à la chaîne de documents avec laquelle sa similarité est la plus grande, en créant une nouvelle chaîne si la similarité est inférieure à un seuil s. Certains auteurs ont tenté ce type d'approches avec des tweets (Petrovic et al., 2010) mais leur modèle inclut tous les tweets collectés aléatoirement par l'API sample de Twitter, y compris ceux traitant de sujets non liés à l'actualité. Étant données les restriction de l'API (accès limité à 1% du flux mondial à un moment t), cette méthode ne peut pas garantir que l'on obtienne tous les tweets liés à un événement donné.

Enfin, un outil mis en place par le Centre Commun de Recherche de la Commission Européenne (Tanev et al., 2012) s'inscrit dans la même démarche que la nôtre puisqu'il vise à collecter en continu des tweets liés à des articles de presse. Les auteurs utilisent pour ce faire un corpus d'un million d'articles pour calculer le poids (tf-idf) à attribuer aux termes de chaque nouvel article, et utilisent ensuite les termes ayant les meilleurs scores pour formuler des requêtes à l'API search de Twitter. L'inconvénient d'une telle méthode est d'utiliser uniquement le vocabulaire de la presse pour interroger Twitter. Or les écarts sont souvent importants entre le vocabulaire des médias traditionnel et celui employé sur le réseau social, du fait de la nature différente de l'information transmise par les tweets (commentaires, opinions personnelles, rumeurs) (Hoang-Vu et al., 2014).

L'apport de notre approche consiste donc à reformuler les premières requêtes envoyées à l'API de Twitter en fonction des tweets obtenus grâce à celles-ci, afin de prendre en compte les spécificités de la communication sur le réseau social (abbréviations, hashtags, langage familier, fautes d'orthographe).

2. Selon un communiqué de presse de Médiamétrie, Facebook compte 35 millions de visiteurs uniques par mois (en comptabilisant uniquement les visites depuis un mobile), contre 16 millions pour Twitter en juillet 2016.

3. Selon une étude réalisée à la demande de la Commission Européenne auprès de 135 journalistes européens, les journalistes établissent une distinction entre Twitter, largement utilisé pour des raisons professionnelles, et Facebook, dont il est davantage fait un usage privé

3 Démarche adoptée

Le prototype mis en place [4], contrairement à l'outil de Tanev et al. (2012), n'intègre pas de connaissances extérieures et se fonde uniquement sur l'analyse des tweets collectés, dans un esprit robuste. Il prend en entrée les dépêches émises par l'AFP, qui constituent donc la représentation des « événements médiatiques » que l'on cherche à étudier. Pour chaque dépêche, on extrait les entités nommées (noms de lieux, de personnes et d'organisations) présentes dans le titre. Ce sont ces entités qui constituent les requêtes initiales envoyées à l'API search de Twitter. Ainsi, pour la dépêche AFP intitulée « Ziad Takieddine affirme avoir remis trois valises d'argent libyen à Nicolas Sarkozy et Claude Guéant », les termes automatiquement extraits sont « Takieddine », « libyen », « Sarkozy » et « Guéant ». La collecte s'effectue ensuite de la manière suivante : on collecte les tweets en français contenant les termes extraits (t1 AND t2 ... AND tn), dans une fenêtre de 24h autour de la date de la dépêche. Cette étape est répétée chaque jour avec les nouvelles dépêches. Pour les dépêches dont le titre ne contient aucune entité nommée, la requête envoyée à Twitter contient tous les mots du titre. Cette méthode permet d'obtenir les tweets contenant un lien url vers la dépêche AFP ou vers un article de presse en ligne ayant le même titre que la dépêche AFP (ce qui est relativement fréquent).

Dans un second temps, on procède à la reformulation des requêtes. On calcule un score tf-idf pour les mots des tweets collectés : on considère que l'ensemble des tweets associés à chaque événement constitue un "document" et que la "collection" est formée par tous les documents créés dans les 30 jours précédents. Ainsi, on peut attribuer un poids à chaque terme de chaque événement. On sélectionne alors les termes ayant un score tf-idf supérieur à certains seuils s1, s2 et s3 fixés manuellement actuellement (car l'évaluation des résultats n'est pas encore formalisée) pour les 1, 2 et 3-grammes. Ce sont ces termes qui sont utilisés pour formuler les prochaines requêtes à l'API. Dans l'exemple proposé ci-dessus, les termes utilisés pour les prochaines requêtes seraient le bigramme « argent lybien » et le hashtag « #Takieddine ».

Cette approche, si elle s'appuie sur le texte des tweets pour améliorer les requêtes, reste cependant très liée au titre des dépêches initiales, du fait de la brièveté des tweets. Le modèle word2vec permet d'identifier de nouveaux termes, différents de ceux présents dans le titre des dépêches mais employés dans un contexte similaire. Word2vec est un modèle prédictif proposé en 2013 permettant d'apprendre un « word embedding » à partir d'une large collection de textes (Mikolov et al., 2013). Il s'agit donc de représenter les mots d'une collection de textes par des vecteurs de réels, les termes employés dans un contexte similaire ayant des vecteurs proches. Cela permet par la suite d'obtenir, pour un terme donné, le ou les termes fréquemment employés dans le même contexte. Nos essais, réalisés en entraînant le modèle sur 2 millions de tweets en français collectés aléatoirement via l'API streaming, montrent qu'il permet effectivement de formuler des synonymes, des abbréviations (« fh » pour « François Hollande ») ou des orthographes alternatives (« Trierweiller » pour « Trierweiler »). Cependant, le modèle renvoie aussi d'autres types de résultats : des termes employés dans un contexte proche mais qui ne sont pas des synonymes. Ainsi, « Juppé », « Macron » ou « Valls » sont également des termes renvoyés par le modèle comme proches du terme « Hollande ». L'utilisation de word2vec nécessite donc un post-process de filtrage des termes pertinents pour les requêtes, fondé sur l'étude de la répartition temporelle des tweets.

4. L'outil se compose d'une partie extraction d'entités nommées, codé avec Knime, et d'une partie extension de requêtes, en Python. Cette partie est disponible sur Github : http://www.github.com/bmaz/quick_twython. L'outil nécessite l'installation d'Elasticsearch. Le modèle word2vec entraîné pour les tests est disponible dans le même dépôt.

4 Conclusion

L'outil mis en place permet la formulation de requêtes à l'API de Twitter en utilisant un vocabulaire spécifique aux usages du réseau social, grâce à l'identification des termes récurrents dans les corpus de tweets collectés. Par ailleurs, la recherche de synonymes via le modèle word2vec, si elle n'a pour l'instant été testée qu'à petite échelle, permet d'envisager la formulation de requêtes indépendantes des dépêches AFP initiales, à condition de mettre en place une étape de filtrage des termes utilisés. À cette fin, nous envisageons un filtre se fondant sur la répartition temporelle des tweets obtenus : une requête pertinente peut être identifiée si elle permet la collecte de tweets très concentrés autour de la date et l'heure de l'événement étudié. L'ajout de ce filtre à notre prototype devrait permettre la formulation de requêtes à la fois plus nombreuses et plus fiables pour collecter des corpus sans bruit.

Références

Hoang-Vu, T.-A., A. Bessa, L. Barbosa, et J. Freire (2014). Bridging vocabularies to link tweets and news. *WebDB*.

Hu, Y., A. John, F. Wangand, D. Duncan-Seligmann, et S. Kambhampati (2012). Et-lda: Joint topic modeling for aligning, analyzing and sensemaking of public events and their twitter feeds. *AAAI*, 59–65.

Hua, T., Y. Ning, F. Chen, C.-T. Lu, et N. Ramakrishnan (2016). Topical analysis of interactions between news and social media. *AAAI*, 2964–2971.

Mikolov, T., K. Chen, G. Corrado, et J. Dean (2013). Efficient estimation of word representations in vector space. *ICLR Workshop*.

Petrovic, S., M. Osborne, et V. Lavrenko (2010). Streaming first story detection with application to twitter. *HLT-NAACL*, 181–189.

Rieder, B. et N. Smyrnaios (2012). Pluralisme et infomédiation sociale de l'actualité : le cas de twitter. *Réseaux 6*, 105–139.

Tanev, H., M. Ehrmann, J. Piskorski, et V. Zavarella (2012). Enhancing event descriptions through twitter mining. *ICWSM*.

Zhao, W. X., J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, et X. Li (2011). Comparing twitter and traditional media using topic models. *ECIR*, 338–349.

Summary

This article is part of a research project focused on studying traditional media and social networks and their mutual influences. Our study aims at creating a tool collecting a continuous stream of tweets related to media events. Our goal is to get both exhaustive and precise (noise minimization) collections of tweets referring to each media event. Our methodology is based on an iterative process : tweets collected at the first step are analyzed in order to improve next collections. Out tool delivers good results concerning the relevance of collected tweets to the media events. However progress could be made to collect larger sets of tweets, that is why we suggest ways in which the tool could be improved.

Analyse exploratoire de corpus textuels pour le journalisme d'investigation

Nicolas Médoc*,** Mohammad Ghoniem**
Mohamed Nadif*

*LIPADE, Université Paris-Descartes
mohamed.nadif@mi.parisdescartes.fr
**Luxembourg Institute of Science and Technology
nicolas.medoc@list.lu,
mohammad.ghoniem@list.lu

Résumé. Nous proposons un outil de visualisation analytique conçu pour et avec une journaliste d'investigation pour l'exploration de corpus textuels. Notre outil combine une technique de biclustering disjoint pour extraire des sujets de haut niveau, avec une méthode de biclustering non-disjoint pour révéler plus finement les variantes de sujets. Une vue d'ensemble des sujets de haut niveau est proposée sous forme d'une treemap, puis une visualisation hiérarchique radiale coordonnée avec une heatmap permet d'inspecter et de comparer les variantes de sujet et d'accéder aux contenus d'origine à la demande.

1 Introduction

Nous présentons un outil de visualisation analytique conçu pour faciliter l'exploration de grand corpus par des journalistes d'investigation. Ces journalistes commencent typiquement par se faire une idée générale du sujet de leur investigation, puis se concentrent sur l'identification de faits et de points de vue qui confirment ou infirment leur hypothèse de travail. Les corpus textuels sont souvent modélisés par des matrices *Termes*×*Documents*, construites avec la pondération *TF-IDF* sur la base des noms et des verbes lemmatisés. On peut en extraire des sujets à l'aide de `Coclus`, une technique de biclustering diagonal basé sur la modularité de graphes (**?**). On a souvent recours aux nuages de mots pour représenter un sujet décrit par un ensemble de termes associés aux documents qui en traitent. Nous les affichons dans une carte pondérée des sujets. Après avoir identifié un sujet d'intérêt, l'attention du journaliste se porte sur la compréhension de ses variantes. Il s'agit de biclusters non-disjoints mettant en relation des sous-ensembles de documents qui partagent des cooccurrences de termes. Ces variantes peuvent révéler des faits, des points de vue ou des angles d'analyse partagés par plusieurs sources. Les biclusters non-disjoints ont été visualisés de différentes manières, e.g. sous la forme d'enveloppes non-disjointes dans des diagrammes nœuds-liens, des vues matricielles et des coordonnées parallèles par **?**. Dans `BiSet`, **?** utilisent des graphes bipartites chaînés avec des regroupements sémantiques pour représenter les relations de chaînage entre les biclusters. Pour fournir une vue d'ensemble claire d'un grand nombre de biclusters non-disjoints, nous

proposons une visualisation hiérarchique radiale qui permet d'identifier les termes qui les rapprochent ou les distinguent.

2 Vue d'ensemble de l'outil

La vue *Weighted Topic Map* de la Figure **??** est une vue hybride combinant une treemap où chaque sujet extrait par `Coclus` est représenté par un rectangle de surface proportionnelle à son importance. Chaque rectangle contient un nuage de mots détaillant les termes du sujet. La taille et la couleur des mots reflètent respectivement leur représentativité (*TF-IDF*) et le nombre de documents où ils apparaissent. Une projection MDS calculée à partir de la matrice de similarité des biclusters de `Coclus` fournit des positions 2D qui servent à placer les rectangles de la vue *Weighted Topic Map* **?**. Ainsi, les sujets similaires se retrouvent dans des rectangles voisins. L'indice de Jaccard est utilisé pour afficher interactivement les liens entre un sujet cible et les cinq sujets les plus similaires. L'affichage de ces liens vise à atténuer les effets du partitionnement strict de `Coclus`. Quand l'analyste sélectionne un sujet pour

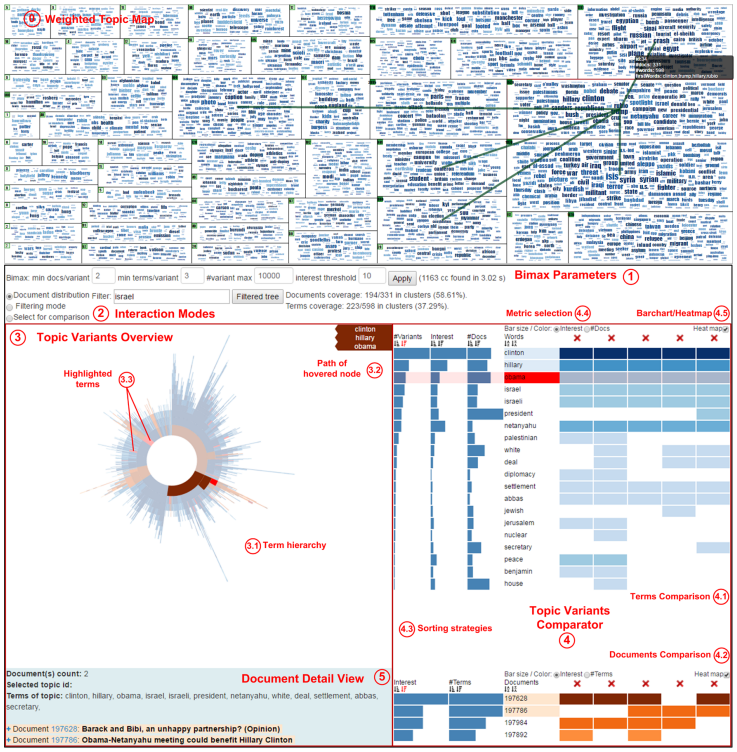


FIG. 1 – *Sujet relatif aux élections présidentielles américaines sélectionné parmi 3 992 articles de presse en ligne compilés entre le 2 et le 16 novembre 2015. Cinq variantes concernant Hillary Clinton sont placés dans le comparateur (*`https://youtu.be/rj9YrTMPClQ`*).*

l'examiner, `Bimax` (**?**), un algorithme de biclustering non-disjoint à base de motifs, en extrait

les variantes. Bien que l'exhaustivité de `Bimax` soit conforme aux besoins de l'analyste, cet algorithme produit une myriade de biclusters. Pour comprendre le résultat de `Bimax`, nous créons une hiérarchie de biclusters sur la base de leurs termes communs à l'aide de l'algorithme `FPTree` (**?**). L'arborescence qui en découle est représentée à l'aide d'une vue *Sunburst* (3.1 dans la Figure **??**). Les termes les plus communs ont un degré de chevauchement plus élevé, et sont placés à proximité de la racine, alors que les termes plus spécifiques sont placés plus en périphérie. Chaque chemin allant de la racine jusqu'à une feuille, décrit les termes d'un bicluster. À mesure que l'on s'éloigne de la racine le long de ce chemin, la combinaison de termes devient plus spécifique et caractérise de moins en moins de documents. Au niveau d'une feuille, on retrouve les documents correspondant à un seul bicluster. À l'aide de cette vue et de la vue *Variant Comparator* (4), le journaliste peut se concentrer sur un aspect spécifique du sujet et afficher les liens entre les documents pertinents, pour identifier les faits et les points de vues relatifs à son hypothèse de travail. Le texte des documents est accessible via la vue *Document Detail* (5). De plus, nous fournissons plusieurs modes d'interaction permettant de filtrer les variantes par mots clés et d'analyser la dispersion de ses documents. En survolant un terme de l'arborescence, toutes ses occurrences sont surlignées en rouge (3.3 dans la Figure **??**) et la séquence de termes correspondante est affichée à droite (3.2). Le comparateur de variantes permet l'analyse des termes communs et distinctifs, ainsi que la distribution des documents à travers les variantes de sujet sélectionnées. Différents critères de tri sont proposés pour faciliter l'identification des termes les plus informatifs.

Le nombre de biclusters `Bimax` augmente avec la taille et la densité des blocs extraits par `Coclus` et peut excéder les 10 000 biclusters. Pour réduire ce nombre, nous permettons à l'utilisateur de modifier les paramètres de `Bimax` : le nombre minimal de termes ou de documents par bicluster (*MinT*, *MinD*) et le nombre maximal de biclusters (*MaxB*). Comme `Bimax` s'applique à des matrices binaires, nous autorisons aussi l'utilisateur à changer le seuil de binarisation (*Thr*) appliqué à la matrice de poids *TF-IDF*. L'augmentation de ce seuil sélectionne, pour chaque document, les termes les plus représentatifs et réduit la densité et les dimensions de la matrice. La Figure **??**, montre l'effet des variations des paramètres sur la hiérarchie de termes associée au sujet concernant les élections présidentielles américaines. Après chaque variation de paramètre, le nœud racine « Obama » est sélectionné systématiquement pour apprécier en orange la distribution de ses documents. Avec les paramètres par défaut ($MinT = 3$, $MinD = 4$, $Thr = 5$), seuls les premiers niveaux des 13 000 biclusters sont visibles dans la *Sunburst*. Augmenter *Thr* ou *MinT* réduit la dispersion des documents concernant « Obama », en préservant mieux la morphologie de la hiérarchie avec *Thr*. Enfin, lorsque *MinD* augmente, la cardinalité des termes des biclusters tend à décroître mais la dispersion des documents sélectionnés demeure jusqu'à ce que le nœud « Obama » disparaisse.

Lors d'une évuation qualitative, nous avons reccueilli les commentaires d'une journaliste d'investigation. D'abord, la carte des sujets a été appréciée. Elle permet de comprendre rapidement des dizaines de sujets de corpus volumineux qu'elle n'est pas capable de traiter sans outil. De plus, « les variantes de sujet couvrent des aspects variés et très précis. » Le comparateur avec ses différents modes de tri « fait gagner du temps » en se focalisant sur les termes les plus pertinents. Certaines limites ont aussi été évoquées comme celles du partitionnement strict du vocabulaire dans les sujets et un besoin de mieux se repérer dans la vue sunburst. Enfin, nous envisageons de mener une étude utilisateur pour évaluer la faisabilité des tâches des journalistes, en comparant les différences entre `Coclus` et `LDA` à travers nos visualisations.

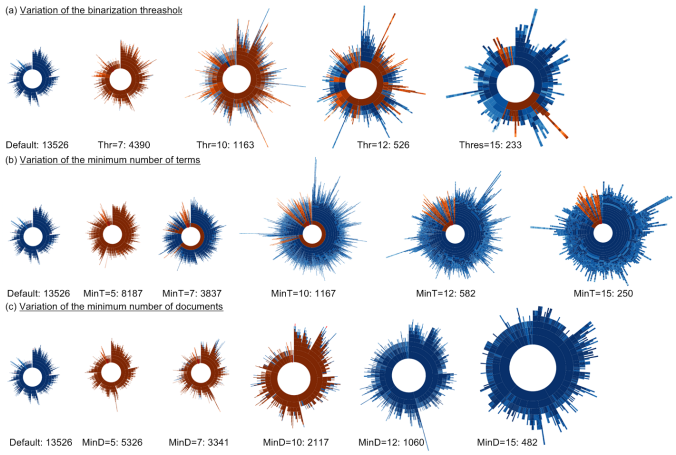


FIG. 2 – *Nombre de biclusters lorsque les paramètres de* `Bimax` *varient.*

Références

Ailem, M., F. Role, et M. Nadif (2015). Co-clustering Document-term Matrices by Direct Maximization of Graph Modularity. In *Proc. of the 24th ACM International on CIKM*, CIKM '15, pp. 1807–1810. ACM.

Ghoniem, M., M. Cornil, B. Broeksema, M. Stefas, et B. Otjacques (2015). Weighted maps : treemap visualization of geolocated quantitative data. In *IS&T/SPIE Electronic Imaging*, pp. 93970G–93970G. Int. Soc. for Optics and Photonics.

Han, J., J. Pei, et Y. Yin (2000). Mining Frequent Patterns Without Candidate Generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pp. 1–12. ACM.

Prelić, A., S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, et E. Zitzler (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics 22*(9), 1122–1129.

Santamaría, R., R. Therón, et L. Quintales (2008). A visual analytics approach for understanding biclustering results from microarray data. *BMC Bioinformatics 9*(1), 247.

Sun, M., P. Mi, C. North, et N. Ramakrishnan (2015). BiSet : Semantic Edge Bundling with Biclusters for Sensemaking. *IEEE TVCG PP*(99), 1–1.

Summary

We propose a visual analytics tool to support investigative journalists in the exploration of large text corpora. Our tool combines graph modularity-based diagonal biclustering to extract high-level topics with overlapping bi-clustering to elicit fine-grained topic variants. Our coordinate and multi-resolution views allows explorin high-level topics, inspecting their variants while accessing the original content on demand.

Détection automatique de grandes thématiques de la propagande Nord Coréenne

Natalia Grabar*, Mason Richey**

* CNRS, UMR 8163, F-59000 Lille, France
Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
natalia.grabar@univ-lille3.fr
http://natalia.grabar.free.fr/
**Graduate School of International and Area Studies
Hankuk University of Foreign Studies, South Korea
mrichey@hufs.ac.kr

Résumé. Les rhétoriques utilisées pour mener les politiques nationale et internationale sont liées entre elles, mais ne montrent pas souvent les mêmes priorités ni les mêmes modes de communication. Lorsque ces deux rhétoriques sont bien utilisées, elles permettent de maximiser la crédibilité et le potentiel de coercition d'un pays, tout en signalant son ouverture vers la coopération. La propagande Nord Coréenne est particulière parce qu'elle se base sur une rhétorique agressive dans sa politique internationale et parce qu'elle aborde des thématiques bien spécifiques. Nous proposons de distinguer automatiquement les grandes thématiques abordées dans la propagande Nord Coréenne en utilisant des scores de similarité lexicale entre deux articles et des méthodes de clustering non supervisé. L'évaluation des clusters obtenus indique que parmi les thématiques les plus présentes se trouvent par exemple: la propagande contre les États-Unis, le Japon et la Corée du Sud; les voeux et félicitations à destination de différents pays, présidents, institutions et personnalités; les honneurs reçus par les dirigeants de la Corée du Nord; les achèvements et avancées du pays et du peuple Nord Coréen; les relations officielles à différents niveaux; les nouvelles du quotidien Rodong Sinmun; les nouvelles inter-coréennes; la culture traditionnelle. Nous avons plusieurs perspectives à ce travail.

1 Introduction

Les politiques nationale et internationale, même si elle sont liées entre elles, ne montrent pas souvent les mêmes priorités ni les mêmes modes de communication. De plus, la rhétorique qui est utilisée avec succès pour l'une peut avoir plutôt un effet contraire sur l'autre. La plupart de dirigeants des pays arrivent à adapter leurs discours pour refléter leur orientation politique. L'exercice n'est pas évident. En voilà deux exemples :

– la politique intérieure est souvent appuyée par le protectionnisme de la nation dans différents domaines (*e.g.*, économie, emploi, sécurité), alors que la politique extérieure gagne

en montrant l'ouverture d'un pays dans différents domaines (*e.g.*, économie, emploi, sécurité) ;
– les politiciens emploient souvent un langage relativement dur en politique intérieure, tandis qu'ils utilisent un langage relativement diplomatique en politique extérieure.

Les théoriciens de relations internationales et les praticiens d'affaires étrangères comprennent très bien les enjeux et la difficulté de concilier les deux rhétoriques. Lorsque ces deux types de discours sont bien utilisés, ils permettent de maximiser la crédibilité et le potentiel de coercition d'un pays, tout en signalant son ouverture vers la coopération. Il existe cependant le danger de la perte de crédibilité d'un dirigeant dû au fait que, après une période d'escalade d'agressivité au niveau international, il revient en arrière et diminue cette agressivité dans son discours. En général, ceci est assez mal perçu au niveau national. C'est ce qu'on appelle *"audience cost"* (Fearon, 1994; Weeks, 2008; Weiss, 2013). La ligne politique doit donc rester bien cohérente.

Le bon dosage des deux modes de communication est important. Lorsqu'il n'est pas respecté, cela peut mener vers l'incompréhension des attentes et des intentions et même vers des situations de crise. La Corée du Nord est un exemple déconcertant de politique extérieure et intérieure, augmenté par les provocations militaires systématiques (Richey, 2015; Westby, 2014). D'une part, la Corée du Nord diffère d'autres pays par le fait que son audience et l'impact interne sont faibles à cause du régime dictatorial. D'autre part, la rhétorique politique est particulière à cause du niveau extraordinaire de grandiloquence et de belligérance, qui se manifestent dans la communication intérieure et extérieure. En effet, sa rhétorique belliqueuse au niveau international est fameuse pour ses locutions extrêmement créatives et hyperboliques :

– *Séoul va être transformé en "une mer de feu"*
– *l'armée de la Corée du Nord est "prête à mener la guerre sainte contre Séoul"*
– *la Corée du Nord va utiliser les armes nucléaires contre la Maison Blanche et le Pentagone, qui sont "les sources du mal"*
– *les États-Unis est le "colonisateur Yankee" et ses citoyens sont des "barbares"*
– *le président Obama est un "proxénète", Lee Myung Bak une "marionnette américaine", alors que le président Park se distingue par son "sifflement venimeux d'une mauviette"*

Ce qui est intéressant est que ces locutions sont systématiquement utilisées pour parler d'autres pays, typiquement des États-Unis, du Japon et de la Corée du Sud, et de leurs dirigeants, et font souvent partie de réponses que la Corée du Nord fait face aux actions politiques éventuelles, comme par exemple la condamnation de la Corée du Nord à cause du non-respect des droits de l'homme, les sanctions de l'ONU ou des entraînements militaires communs des États-Unis et de la Corée du Sud. Par ailleurs, même si la plupart de ces locutions sont créées en coréen, elles sont souvent traduites en anglais par la chaîne KCNA [1] pour être ensuite diffusées à l'international. Comme la grande majorité de Nord-Coréens ne parle pas anglais, on peut supposer que ces traductions sont vraiment destinées pour l'audience internationale.

Malgré la nature intéressante de la politique et du discours nord coréens, il existe très peu de travaux qui y sont consacrés. Nous proposons de contribuer à l'étude de la propagande politique, sur l'exemple de la propagande produite par la Corée du Nord. Dans la suite de cette contribution, nous présentons d'abord des travaux existants en relation avec le sujet traité (section 2), nous décrivons ensuite la méthode utilisée (section 3). Nous présentons et discutons les résultats (section 5) et concluons avec quelques perspectives à ce travail (section 6).

1. `http://www.kcna.us/`

2 Travaux existants

Les écrits journalistiques et politiques sont objet de plusieurs travaux en TAL, lexicométrie, étude discursive et étude de corpus de manière générale. Mentionnons par exemple : l'analyse générale du discours syndical (Habert, 1983; Salem, 1993) et politique (Salem, 1981), la propagation des informations sur les réseaux (Bourigault et al., 2014), la détection de buzz et de fausses rumeurs (Chou et al., 2015; Ma et al., 2015; Fuchs et Yu, 2015), la véracité des informations et les nouveaux modèles du journalisme sur les réseaux (Derczynski et Bontcheva, 2014; Sharma, 2015; Maigrot et al., 2016).

En relation avec la Corée du Nord, les chercheurs se concentrent sur différentes thématiques : le discours politique nord coréen et son impact sur la sécurité locale et mondiale (Myers, 2010, 2015; Richey, 2015; Ohn et Richey, 2015), le programme nucléaire nord coréen (Rich, 2012, 2014b), la rhétorique belliqueuse, essentiellement en relation avec la provocation militaire (Joo, 2015), l'émergence médiatique de leaders (Rich, 2014a), comparaison de discours de Kim Il Sung et Fidel Castro (Malici et Malici, 2005). Par ailleurs, les unités des analyses plus généralistes peuvent être les mois, les semaines (Zuell, 2010) ou les jours (Rich, 2012), avec des données qui s'étendent sur une à trois années. Dans la plupart de travaux existants, l'analyse du discours est effectuée manuellement par les chercheurs qui viennent essentiellement des domaines de relations internationales et d'études politiques. Dans de rares cas où des méthodes automatiques sont utilisées, elles exploitent : (1) des techniques de data mining et de lexicométrie, comme par exemple les pondérations (fréquences, tfidf ou jaccard) ou la régression binomiale d'un ensemble de termes prédéfinis (Haynes, 2001; Rich et Liu, 2012) ; (2) l'apprentissage supervisé pour détecter les articles provocatifs de la Corée du Nord. Ainsi, avec cinq mots-clés (*years, signed, assembly, June, Japanese*) une précision de 82 % est obtenue (Whang et al., 2016). Il existe également des travaux qui étudient la propagande et le discours extrémiste produit par d'autres pays ou groupes : détection automatique de commentaires commandités par le gouvernement de Chine (Blake et Miller, 2016) et détection de traces digitales d'extrémistes solitaires sur les réseaux (Chen, 2007; Brynielsson et al., 2012).

L'objectif de notre travail consiste à analyser les articles de l'agence de presse KCNA de la Corée du Nord pour détecter les grandes thématiques des articles de propagande. Nous abordons cette question de recherche comme un problème de clustering de textes.

3 Méthodes

Nous présentons d'abord les données étudiées et ensuite les différentes étapes de la méthode : pré-traitement, calcul de similarité entre les articles, le clustering et l'évaluation.

3.1 Données de la propagande Nord Coréenne

Nous avons collecté les articles du site KCNA qui fournit la propagande officielle de la Corée du Nord. Créée en 1946, l'agence KCNA effectue une publication journalière en ligne d'articles en anglais depuis 1997. Ce site ne propose que des données textuelles (pas d'images ni de vidéos). Il est admis que l'utilisation de la langue anglaise pour la propagande de KCNA cible une audience différente par rapport à la propagande en langue coréenne et conduit donc à

une différence de contenu aussi (Poneman et al., 2004). Néanmoins, l'exploitation de la propagande produite et traduite directement par une agence Nord Coréenne pour le public étranger permet d'éviter les biais de perception occidentale de ce type de littérature idéologique. De plus, les données sont accessibles en ligne pour plusieurs années maintenant.

Le corpus total contient 121 964 articles, soit 31 211 998 occurrence de mots. Le volume de la propagande va en augmentant au fil des années, autant en nombre d'articles que d'occurrences de mots. Le pic des années 2011 et 2012 est sans doute causé par les changements de leaders politiques dans le pays : l'émergence médiatique et politique de Kim Jong-Un, suite à la maladie de son père Kim Jong-Il et surtout par rapport à ses deux frères aînés, et la transition entre le père Kim Jong-Il et son fils cadet Kim Jong-Un. Nous proposons de nous concentrer sur deux années : 2003 (4 852 articles, 1 302 220 occ.) et 2013 (9 967 articles, 2 766 178 occ.). La méthode, ajustée sur ces données, pourra ensuite être testée sur le corpus entier.

3.2 Pré-traitement du corpus

Nous effectuons une série de pré-traitement du corpus :

- la conversion du format HTML au format texte ;
- la séparation des articles selon la langue dans laquelle ils sont écrits. Nous exploitons les listes de mots proposées dans un travail précédents pour distinguer l'anglais, le français et l'allemand (Grefenstette et Nioche, 2000) et y ajoutons des mots fréquents et typiques pour distinguer l'espagnol ;
- la suppression de motifs systématiquement insérés dans les articles, comme par exemple la date ou l'année de Juche, qui est l'idéologie autocratique développée par le 1er président de la Corée du Nord Kim Il-Sung et sur laquelle repose le régime de la Corée du Nord. À titre d'information, 2017 est l'année Juche 107 ;
- la création de trois ensembles de données : articles complets, titres des articles et corps des articles ;
- l'étiquetage morpho-syntaxique avec Treetagger (Schmid, 1994).

3.3 Calcul de similarité

La similarité entre chaque paire d'articles est calculée avec `Text::Similarity` [2], un module écrit en Perl et permettant de calculer l'intersection lexicale entre les textes traités. Ce module effectue des traitements supplémentaires : la suppression de la ponctuation, la minusculisation et la suppression de mots grammaticaux. La similarité est estimée être le nombre de mots communs dans les deux fichiers comparés, pondéré par la longueur de chaque fichier. Les valeurs de similarité sont entre 0 et 1.

3.4 Clustering

Le clustering est effectué avec l'algorithme `MCL` (Markov Cluster Algorithm), qui permet d'effectuer un clustering non supervisé sur des graphes (van Dongen, 2000). Cet algorthme a plusieurs avantages importants pour nous : il est simple d'utilisation, très rapide et il n'est pas nécessaire de lui indiquer le nombre de clusters attendus en sortie. En revanche, il est

2. `http://search.cpan.org/dist/Text-Similarity/lib/Text/Similarity.pm`

possible de modifier la valeur d'un des paramètre `-I`, qui permet de régler la granularité des clusters : plus cette valeur est élevée plus la granularité des clusters est fine. Cet algorithme a été exploité avec de différents types de données et nous proposons de le tester sur les données de la propagande.

3.5 Évaluation

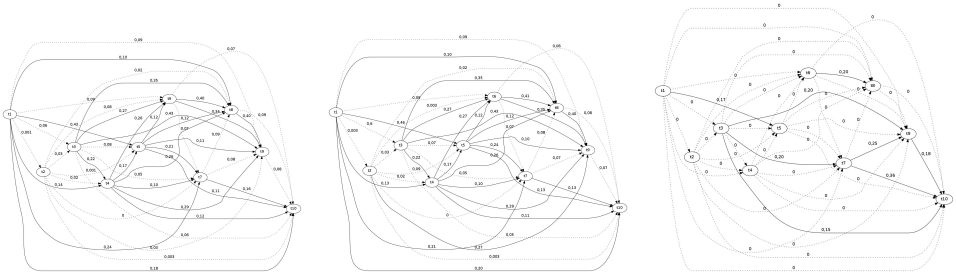
L'évaluation est effectuée manuellement et *a posteriori* des traitements automatiques. L'objectif est d'analyser le contenu des clusters obtenus. Pour ceci, nous analysons : (1) les mots les plus fréquents de chaque cluster et (2) les titres des articles de chaque cluster. Étant donné que les mots fréquents et les titres sont assez explicites, cela permet d'avoir un premier jugement sur la thématique des clusters et leur homogénéité.

4 Rationale de l'étude

Nous effectuons plusieurs tests, en variant plusieurs paramètres. Pour deux types d'unités linguistiques (formes et lemmes), nous étudions différents types d'unités textuelles : les articles entiers, les titres des articles, le corps des articles (sans leurs titres). Le recouvrement lexical entre deux unités textuelles est retenu comme valeur de similarité. Lors du clustering, nous testons plusieurs valeurs du paramètre `-I`, qui influence la granularité des clusters, dans l'intervalle [2,0, 9,5], en l'incrémentant par 0,5 point. Notons que 5,0 est la valeur par défaut de `I`. Les résultats présentés concernent les traitements effectués sur deux années, avec 10 ans de différence : 2003 et 2013. Le travail est effectué sur les articles écrits en anglais.

5 Résultats et discussion

Après la reconnaissance et le filtrage de la langue, nous retenons 3 926 articles en 2003 et 8 472 articles en 2013. D'autres articles de ces deux années sont soit en espagnol soit des articles trop courts pour lesquels la décision sur la langue ne peut pas être faite.



(a) Articles entiers (b) Corps des articles (c) Titres des articles

FIG. 1 – *Un exemple des valeurs de similarité au sein d'un ensemble de 10 articles.*

Nous supposons que les différents paramètres indiqués dans la section 4 vont avoir une influence sur les résultats. Par exemple, la séparation des articles en titre et corps de l'article et le calcul de similarité sur ces unités textuelles permet d'obtenir des valeurs de similarité différentes, comme présenté dans la figure 1 pour un ensemble de 10 articles. Sur cette figure, les noeuds représentent les articles alors que les arcs représentent les valeurs de similarité calculées entre chaque paire d'articles. Nous avons mis en pointillé les arcs dont les valeurs de similarité sont nulles ou inférieures à 0,10. Nous pouvons voir que les graphes obtenus avec les articles entiers et le corps des articles ont une forme similaire, avec toutefois des valeurs de similarité un peu différentes. Le graphe obtenu avec les titres est d'une part plus *"sélectif"* car il comporte beaucoup plus de valeurs de similarité nulles, et d'autre part il est différent car les similarités plus fortes ne relient pas toujours les mêmes articles. L'utilisation des formes et des lemmes produit les valeurs de similarité de mêmes grandeurs. Pour certaines paires d'articles, la similarité basée sur les lemmes est un peu plus élevée que celle obtenue avec les formes car les lemmes, en permettant de regrouper certaines formes ensemble, permettent également d'augmenter la similarité entre les textes.

`I`	*2003, Titres*	*2013, Titres*
4,0	99	70
4,5	118	83
5,0	143	107
5,5	176	121
6,0	196	145
6,5	220	165
7,0	240	181
7,5	266	199
8,0	292	222
8,5	316	246
9,0	344	263
9,5	363	276

TAB. 1 – *Nombre de clusters selon le paramètre `I` de `MCL`, avec les titres et les lemmes.*

Lorsque les graphes de similarités sont calculés, ils peuvent être exploités pour générer les clusters. Dans le tableau 1, nous présentons le nombre de clusters obtenus avec les valeurs de `I` allant de 4,0 à 9,5. Dans les tableaux 2 et 3, nous présentons les cinq clusters les plus gros, générés avec les titres des articles, les lemmes et `I`=5,0. Nous indiquons la taille de ces clusters et les mots-clés les plus fréquents. Par exemple, le 1er et 3eme clusters de 2003 et le 1er cluster de 2013 contiennent des articles dirigés contre les États-Unis, le Japon et la Corée du Sud. Les autres clusters de ces tableaux concernent la glorification de la Corée du Nord et de ses leaders à travers des commémorations, des cadeaux reçus et envoyé, des réunions, etc. À côté des clusters très et moyennement grands, nous avons 118 et 93 clusters de moins de 10 articles générés à partir des années 2003 et 2013, respectivement. Sur la base des informations comme celles présentées dans les tableaux 2 et 3, nous pouvons faire émerger quelques thématiques principales (dans l'ordre de leur importance) :

Taille	*Mots-clés*
1 156	*s./324 korea/307 dprk/217 korean/179 u.s./140 call/111 hold/61 people/58 national/50 struggle/44 japan/44 kcna/42 meeting/39 reunification/38 war/38 dispatch/36 troop/36 held/35 foreign/34 anti-u.s/30*
758	*greeting/193 president/169 minister/75 japan/64 dprk/62 prime/49 message/42 urged/38 foreign/32 koreans/32 anniversary/32 call/31 urge/30 korean/26 congratulation/25 envoy/24 special/24 national/23 sympathy/21 hold/20*
693	*u.s./408 dprk/220 fire/104 kcna/101 japan/66 urge/60 korean/47 move/47 s./46 war/42 policy/39 anti-dprk/39 talk/34 nuclear/33 military/30 blast/28 hostile/26 remark/25 urged/24 korea/24*
508	*kim/510 il/434 jong/358 sung/89 yong/60 nam/54 gift/54 kpa/43 floral/41 work/40 basket/39 president/36 unit/36 anecdote/29 inspect/29 anniversary/26 message/26 congratulatory/24 letter/23 publish/22*
84	*reception/60 ambassador/42 give/40 russian/16 host/15 chinese/8 hosts/7 military/7 dprk/7 iranian/6 performance/6 egyptian/4 embassy/4 palestinian/4 attache/4 general/3 cuban/3 participant/2 libyan/2 guest/2*

TAB. 2 – *Exemple de 5 clusters les plus gros en 2003, titres des articles, lemmes, I=5,0.*

1. *Propagande contre les États-Unis, le Japon et la Corée du Sud*. Si les armes, le feu ou la guerre nucléaire y sont très présents, d'autres sujets peuvent également provoquer l'écriture de ces articles, comme la politique étrangère, l'impérialisme, diverses sanctions pas forcément contre la Corée du Nord, des entraînements militaires, l'espionnage contre la Corée du Nord ou les droits de l'homme ;
2. *Voeux et félicitations à destination de différents pays, présidents, institutions et personnalités* (Fidel Castro, Yasser Arafat, President of Sudan, Slovenian President, President of Serbia and Montenegro, President of Trinidad and Tobago, Iranian President, Tunisian President, Tunisian Prime Minister, Syrian President, Uzbek Foreign Minister, Cyprian Foreign Minister, Russian President, Russian Prime Minister...) ;
3. *Glorification des dirigeants de la Corée du Nord* (Kim Il Sung, Kim Jong Il, Pak Pong Ju et ensuite Kim Jong Un). Il s'agit de messages, de félicitations, de lettres, de cadeaux, de fleurs, de cartes reçus mais aussi des oeuvres de ces dirigeants publiées à l'étranger. Notons qu'en 2003, il y a une série d'articles avec des anecdotes sur Kim Il Sung et Kim Jong Il, alors qu'en 2013 ce sujet est absent ;
4. *Achèvements et avancées du pays et du peuple Nord Coréen* dans différents domaines :
   - la nutrition (emballage sous vide du kimchi, nouveau ferment pour le kimchi, céréales riches en gras, thé nutritif, nouvelles variétés de pommiers, de concombres, etc.),
   - la santé (micromanipulateur biologique, peintures anticeptiques, purificateur de sang, produits contre différentes maladies, vaccins pour les animaux, etc.),
   - l'idéologie (timbres, posters, slogans, ouvrages, sites de propagande),
   - l'éducation (programmes et jeux informatiques, programmes éducatifs, dictionnaires),
   - l'industrie (tannage, station de marée motrice, broderie, brassage de bière...) ;
5. *Relations avec les ambassades et ambassadeurs* accompagnées d'événements sociaux ;

Taille	*Mots-clés*
7 147	*kim/1702 dprk/1624 korean/1369 s./1358 jong/1227 il/1051 korea/634 un/600 u.s./564 sinmun/512 rodong/512 war/437 sung/393 people/366 held/347 day/343 anniversary/342 foreign/296 party/278 organization/268*
610	*kim/302 jong/262 un/262 president/214 greeting/186 message/57 floral/51 congratulation/50 gift/43 receives/42 yong/40 basket/40 party/40 congratulatory/38 nam/37 sends/33 leader/32 political/24 letter/24 pm/21 russian/21*
104	*delegation/96 leaves/59 dprk/45 meets/19 government/13 returns/12 china/10 wpk/10 chinese/9 choe/6 mongolian/6 home/6 back/5 russia/5 visit/5 president/5 spa/5 hae/4 friendship/4 ryong/4*
104	*delegation/50 foreign/39 guests/13 chinese/12 arrives/12 leave/11 party/10 government/9 delegations/8 delegates/7 arrive/7 meet/6 indonesian/5 ministry/5 crewmen/4 president/4 meets/3 delegate/3 iiji/3 praise/3*
74	*exhibition/40 opens/40 national/29 held/20 art/14 contest/10 technological/8 scientific/8 photo/7 championship/5 fine/5 achievements/4 intl/4 sports/4 festival/4 song/3 martial/3 trade/3 pyongyang/3 presentation/3*

TAB. 3 – *Exemple de 5 clusters les plus gros en 2013, titres des articles, lemmes, I=5,0.*

6. *Réunions et rassemblements*, y compris Nord-Sud et y compris entre les scientifiques ;
7. *Réunions officielles*, y compris avec la signature de protocoles, accords et traités ;
8. *Nouvelles du quotidien Rodong Sinmun* (*Journal des Travailleurs*), qui est l'organe officiel du Parti du travail de Corée et le journal le plus lu dans le pays ;
9. *Aide aux fermiers* nord coréens venue de l'étranger et acceptée par les leaders du pays ;
10. *Visites des délégations de la Corée du Nord* en étranger ;
11. *Nouvelles inter-coréennes* (groupes de contact, familles séparées, chemins de fer...) ;
12. *Culture traditionnelle* (chants, légendes...).

Avec plusieurs clusters (par exemple, 2, 3, 5, 6, 7 et 10), il est possible de créer un réseau de pays vus comme amicaux par rapport à la Corée du Nord. En fonction de la symétrie des actions et des honneurs, il serait également possible d'établir une échelle de ces amitiés.

Notons que la plupart de ces thématiques sont bien différenciées au sein de clusters dédiés, même s'il est possible d'avoir des intrus dans ces clusters. D'autres thématiques (visites officielles, réunions, commémorations...) sont distribuées entre plusieurs clusters. En ce qui concerne le paramètre `I`, sa valeur par défaut 5,0 semble être optimale dans la génération non supervisée de clusters. Elle offre des clusters en un nombre raisonnable et assez homogènes quant à leur contenu. Nous pensons cependant qu'un `I` plus grand permet de générer des clusters plus fins et homogènes, mais aussi plus distribués. Ceci sera analysé dans un travail futur.

Cette analyse nous donne un aperçu de thématiques principales abordées par la propagande Nord Coréenne. Ces thématiques sont assez stables entre les deux années analysées. Il serait intéressant de comparer ces thématiques avec les sujets abordés dans la presse occidentale. Par exemple, l'économie semble être un des sujets manquants : elle n'est quasiment jamais mentionnée dans le corpus étudié, alors qu'elle occupe une place prépondérante dans la presse

mondiale. Une des raisons est sans doute que, selon la ligne officielle de la Corée du Nord, l'économie nord coréenne, telle que fondée au début de l'existence du pays, est parfaite. Il n'est donc pas nécessaire de discuter ce système économique ou d'essayer de l'améliorer.

Même s'il est difficile d'avoir un aperçu direct et précis de la politique et de la propagande interne de la Corée du Nord, la traduction de cette propagande par les organes de propagande officielle peut néanmoins donner une idée assez claire des lignes principales et privilégiées suivies. De plus, les articles sont écrits de manière très claire et explicite. Comme déjà indiqué dans d'autres travaux (Kim, 1998; Hachten, 1999), nous pensons également que la traduction locale est plus fiable.

6 Conclusion et Perspectives

Nous avons proposé un travail sur la distinction automatique de grandes thématiques de la propagande nord coréenne. Nous utilisons pour ceci un ensemble d'articles collectés sur un site de propagande locale. Nous abordons cette question de recherche comme le problème de clusterisation non supervisée. Le travail est effectué avec les articles en anglais. D'abord, nous calculons la similarité entre les articles par leur recouvrement lexical et ensuite nous effectuons un clustering. Nous travaillons essentiellement avec les titres des articles. L'évaluation est effectuée manuellement et *a posteriori*, en analysant le contenu des clusters obtenus (les mots-clés les plus fréquents et les titres).

Nos résultats permettent d'émerger les thématiques abordées dans les articles de l'agence de presse KCNA. Parmi les thématiques principales se trouvent par exemple : la propagande contre les États-Unis, le Japon et la Corée du Sud ; les voeux et félicitations à destination de différents pays, présidents, institutions et personnalités ; la glorification des dirigeants de la Corée du Nord ; les achèvements et avancées du pays et du peuple Nord Coréen ; les relations officielles à différents niveaux ; les nouvelles du quotidien Rodong Sinmun ; les nouvelles inter-coréennes ; la culture traditionnelle.

Nous avons plusieurs perspectives à ce travail. La mesure de similarité utilisée actuellement est très simpliste : elle calcule seulement le recouvrement lexical. Nous allons exploiter des mesures plus sophistiquées et robustes (Dice, Jaccard, Word2Vec). Nous pensons que cela nous permettra également de raffiner les clusters obtenus actuellement et de travailler sur les articles complets.

Nous traitons actuellement deux années seulement, 2003 et 2013. La méthode proposée pourra être reglée sur cet sous-ensemble et pourra ensuite être appliquée à l'ensemble du corpus (1997 à 2015). Nous pensons obtenir ainsi des clusters plus complets, de faire une comparaison plus complète entre les années et de statuer sur la stabilité des thématiques de la propagande.

D'autres perspectives concernent la comparaison des thématiques abordées dans la propagande nord coréenne avec les sujets abordés dans la presse mondiale. Par ailleurs, il peut être très intéressant de comparer la propagande nord coréenne avec d'autres propagandes, comme par exemple la propagande russe, qui devient de plus en plus offensive et agressive. Comme le montre la figure 2, nous nous attendons à ce que la similarité se manifeste non seulement au niveau textuel, mais également au niveau des représentations.

(a) Kim Il Sung (b) Lénine

FIG. 2 – *Exemple de deux figures politiques en Corée du Nord et en Union Soviétique/Russie*

Remerciements

Ce projet est effectué dans le cadre de l'appel *Projet Partenarial* de la MESHS (Maison Européenne des Sciences de L'homme et de la Société) en Hauts-de-France. Nous remercions également Vincent Claveau et Thierry Hamon pour leurs conseils méthodologiques et le soutien logistique.

Références

Blake, A. et P. Miller (2016). Automated detection of chinese government astroturfers using network and social metadata. *SSRN's eLibrary*.

Bourigault, S., C. Lagnier, S. Lamprier, L. Denoyer, et P. Gallinari (2014). Learning social network embeddings for predicting information diffusion. In ACM (Ed.), *International Conference on Web Search and Data Mining*, New York, NY, USA, pp. 393–402.

Brynielsson, J., A. Horndahl, et F. Johansson (2012). Analysis of weak signals for detecting lone wolf terrorists. In *Intelligence and Security Informatics Conference (EISIC)*.

Chen, H. (2007). Exploring extremism and terrorism on the web : The Dark Web project. *Intelligence and Security Informatics 4430*, 1–20.

Chou, H.-I., G. Y. Tian, et X. Yin (2015). Takeover rumors : Returns and pricing of rumored targets. *International Review of Financial Analysis 41*, 13–27.

Derczynski, L. et K. Bontcheva (2014). Pheme : Veracity in digital social networks. In *Workshop on Interoperable Semantic Annotation (ISA)*.

Fearon, J. (1994). Domestic political audiences and the escalation of international disputes. *American Political Science Review 88*(3), 577–592.

Fuchs, M. et P.-D. Yu (2015). Rumor source detection for rumor spreading on random increasing trees. *Electron. Commun. Probab 20*(2), 1–12.

Grefenstette, G. et J. Nioche (2000). Estimation of English and non-English language use on the WWW. In *Recherche d'Information Assistée par Ordinateur (RIAO)*, Paris, pp. 237–246.

Habert, B. (1983). Études des formes spécifiques et typologie des énoncés (les résolutions générales des congrès de la CFTC-CFDT de 1945 à 1979). *MOTS, Presses de la Fondation Nationale des Sciences Politiques* (7), 97–124.

Hachten, W. (1999). *The World News Prism*. Ames, IA : Iowa State University Press.

Haynes, J. (2001). Red journalism as a keyhole : Evaluating discrepancies in news systems and inferring the political direction of North Korea based on a quantitative content analysis of the Korean central news agency website. Technical report, University of North Carolina. Master's Thesis.

Joo, H.-M. (2015). Predicting North Korean military provocations : Document classification analysis of KCNA news. In *ISA Conference 2015*, West Pasadena.

Kim, S. (1998). *North Korean Foreign Relations in the Post-Cold War Era*. New York : Oxford University Press.

Ma, J., W. Gao, Z. Wei, Y. Lu, et K.-F. Wong (2015). Detect rumors using time series of social context information on microblogging websites. In ACM (Ed.), *CIKM'15 Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, Melbourne, Australia, pp. 1751–1754.

Maigrot, C., E. Kijak, et V. Claveau (2016). Médias traditionnels, médias sociaux : caractériser la réinformation. In *TALN 2016*, pp. 1–8.

Malici, A. et J. Malici (2005). The operational codes of Fidel Castro and Kim Il Sung : the last cold warriors ? *Polit. Psychol 26*(3), 387–412.

Myers, B. (2010). *The Cleanest Race : How North Koreans See Themselves and Why It Matters*. Hoboken, NJ : Melville House.

Myers, B. (2015). *North Korea's Juche Myth*. Busan : Sthele Press.

Ohn, D. et M. Richey (2015). China's evolving policy towards the Democratic People's Republic of Korea under the Xi Jinping leadership. *Asian Studies Review 39*(3).

Poneman, B., J. Wit, et R. Galluci (2004). *Going Critical : the First North Korean Nuclear Crisis*. Washington, DC. : Brookings Institution.

Rich, T. (2012). Deciphering North Korea's nuclear rhetoric : An automated content analysis of KCNA news. *Asian Affairs : An American Review 39*(2), 73–89.

Rich, T. (2014a). Introducing the great successor : North Korean english language news coverage of Kim Jong Un 2010-2011. *Communist and Post-Communist Studies 47*, 127–136.

Rich, T. (2014b). Propaganda with purpose : uncovering patterns in North Korean nuclear coverage, 1997-2012. *International Relations of the Asia-Pacific 14*(3), 427–453.

Rich, T. et T. Liu (2012). Reading between the lines : automated content analysis of North Korean nuclear rhetoric. *Rev. Glob. Polit 38*, 157–176.

Richey, M. (2015). Considering DPRK regime collapse : Its probability and possible geopolitical and security consequences. In *Egmont Security Policy Brief*.

Salem, A. (1981). Signalement et inventaire lexical : textes politiques français de 1793. In *Pratique de l'analyse des données*, pp. 183–197. Paris : Dunod.

Salem, A. (1993). De travailleurs à salariés. Repères pour une étude de l'évolution du vocabulaire syndical. *Mots 36*, 74–83.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *ICNMLP*, Manchester, UK, pp. 44–49.

Sharma, N. (2015). *Facebook journalism : An exploratory study into the news values and role of journalists on Facebook*. Thèse de doctorat, Indiana University, Indiana, USA.

van Dongen, S. (2000). *Graph Clustering by Flow Simulation*. Thèse de doctorat, University of Utrecht, Utrecht, The Netherlands. http ://www.library.uu.nl/digiarchief/dip/diss/1895620/inhoud.htm.

Weeks, J. (2008). Autocratic audience costs : Regime type and signaling resolve. *International Organization 62*, 35–64.

Weiss, J. (2013). Authoritarian signaling, mass audiences, and nationalist protest in china. *International Organization 67*, 1–35.

Westby, T. (2014). *North Korean Nuclear Deterrence : A Myth or a Reality ? An Analysis of North Korean Deterrence Credibility toward the United States and South Korea*. Master thesis, Institutt for Statsvitenskap, Oslo, Norway.

Whang, T., M. Lammbrau, et H. min Joo (2016). Detecting patterns in north korean military provocations : what machine-learning tells us. *Int Relat Asia Pac*.

Zuell, C. (2010). Using computer-assisted text analysis to identify media reported events. In *10th International Conference on Statistical Analysis of Textual Data*.

Summary

Rhetorics used for national and international politics are interlinked, although they do not have the same priorities nor they show the same modes of communication. When these two rhetorics are well used, they allow to maximize the credibility and the coertion potential of the countries, and to signal its opening to the cooperation. The Nord Korean propaganda is particular because it provides agressive rhetorics in its international politics and addresses very specific topics. We propose to distinguish automatically main thematics exploited by the North Korean propaganda using lexical similarity scores between each pair of articles and non-supervised clustering methods. The evaluation of the clusters obtained indicates that between the main thematics we can find for instance: propaganda against United States, Japan and South Korea; greetings and felicitations for different countries, presidents, institutions and persons; honnors received by North Korean leaders; realizations and achievements of the country and people; official relations at different levels; news on and from the Rodong Sinmun journal; inter-korean news; traditionnal culture. We have several directions for future work.

Utilisation conjointe LDA et Word2Vec dans un contexte d'investigation numérique

Julien Maitre*,** Michel Menard*,***
Guillaume Chiron*,**** Alain Bouju*,♯

*L3I, Université de La Rochelle, Avenue Michel Crépeau 17042 La Rochelle
**julien.maitre@univ-lr.fr, ***michel.menard@univ-lr.fr
****guillaume.chiron@univ-lr.fr, ♯alain.bouju@univ-lr.fr

Résumé. L'étude présentée dans cet article s'inscrit dans le contexte du développement d'une plateforme d'analyse automatique de documents associée à un service caché lanceurs d'alerte focalisé sur la révélation de faits/événements/actions en lien avec des problématiques environnementales. Dans le but de traiter de manière automatique les documents textuels révélés par un lanceur d'alerte et portant sur un ou plusieurs faits relatifs à un événement déclencheur, nous proposons de développer un framework d'investigation qui doit répondre au besoin qu'ont les journalistes/politiques/juristes de se munir d'outils d'investigation (extraction, vérification, corrélation) et de représentation de l'information (synthèse, aide à la décision). Il a pour but de faciliter les expertises indépendantes, protéger les lanceurs d'alerte et aider à la détection des signaux faibles. Cet article se focalise plus particulièrement sur le clustering thématique multi-niveaux de documents et l'extraction des indicateurs caractéristiques et significatifs des thèmes. Nous étudions notamment la pertinence d'évaluer une approche s'appuyant sur du comptage de mots par une méthode récente de type "word embedding", *word2vec*. Nous proposons d'évaluer les partitions obtenues grâce à un indice de cohérence sur la collection de mots représentative de chaque thème obtenu. Deux algorithmes sont proposés. Le premier estime le nombre de thèmes le plus pertinent, et extrait ainsi sur ce niveau la collection de mots pour chacun des thèmes trouvés. Le second propose d'extraire les meilleurs collections de mots potentiellement présentes sur des niveaux différents.

1 Introduction

Une problématique majeure actuelle porte sur notre capacité à prendre des décisions éclairées devant l'augmentation drastique des signaux délivrés par toujours plus de moyens d'information. Des phénomènes de saturation des capacités de nos systèmes de traitement conduisent à des difficultés d'interprétation ou même à refuser les signaux précurseurs de faits ou d'événements. L'utilité de la prise de décision contrainte par des nécessités temporelles oblige un traitement rapide de la masse d'information. Etre capable de détecter dans un délai imposé, les bons signaux porteurs de l'information utile dans un contexte de stratégie d'anticipation,

s'avère être un challenge devenu permanent pour de nombreux acteurs économiques. Il est donc nécessaire de développer, sous la forme de plateformes d'investigation (cf. Figure 1), de nouveaux services d'aide à la décision pour les politiques et les organisations en charge de ces activités. Les prises de décision, qui doivent portées aussi bien sur la crédibilité de la source d'information que sur la pertinence des informations révélées relatives à un événement, nécessitent des algorithmes robustes de détection des signaux faibles, d'extraction et d'analyse de l'information portée par ces derniers, d'ouverture sur un contexte informationnel plus large.
Nous proposons de porter notre action sur deux points essentiels : la détection des signaux faibles et l'extraction de l'information véhiculée par ces derniers. Notre objectif concerne donc la détection de signaux précurseurs dont la présence attenante dans un espace de temps et de lieux donnés anticipe l'avènement d'un fait observable. Cette détection est facilitée par l'information précoce délivrée par un lanceur d'alerte sous la forme de documents. Ils exposent des faits avérés, unitaires et ciblés, mais aussi partiaux, relatifs à un événement déclencheur. Le lanceur d'alerte délivre une information non encore décelable/apparente sur les réseaux sociaux et spécialisés. Elle permet de dessiner le contour des signaux faibles à venir sur les réseaux, facilitant ainsi leur détection et l'extraction de l'information portée par ceux-ci.

FIG. 1 – *Plateforme d'investigation. Elle doit répondre au besoin réel qu'ont les journalistes/politiques/juristes de se munir d'outils d'investigation (extraction, vérification, corrélation) et de représentation de l'information (synthèse, aide à la décision). Elle a donc pour but de faciliter les expertises indépendantes, protéger les lanceurs d'alerte et aider à la détection des signaux faibles.*

La procédure d'investigation proposée repose donc sur la détection des signaux faibles présents sur les réseaux. Elle combine algorithmes de fouille de données et visualisation analytique. Elle est facilitée par la connaissance des patterns révélés par le lanceur d'alerte. L'in-

formation est estimée à partir des indicateurs révélés par le lanceur d'alerte et des données portées par les signaux faibles (cf. Figure 2). Les smart data, révélées par le lanceur d'alerte, permettent de mieux cibler le data mining lors des phases de détection des signaux faibles et d'exploration sur les réseaux. Pour le développement du framework d'investigation, trois actions sont donc entreprises :

— Action 1 : Analyse automatique de contenus avec un minimum d'*a priori*. Identification des informations pertinentes. Indicateur de cohérence des thèmes obtenus ;
— Action 2 : Agrégation de connaissances. Enrichissement de l'information. Détection des signaux faibles ;
— Action 3 : Visualisation analytique. Mise en perspective de l'information par la création de représentations visuelles et de tableaux de bord dynamique.

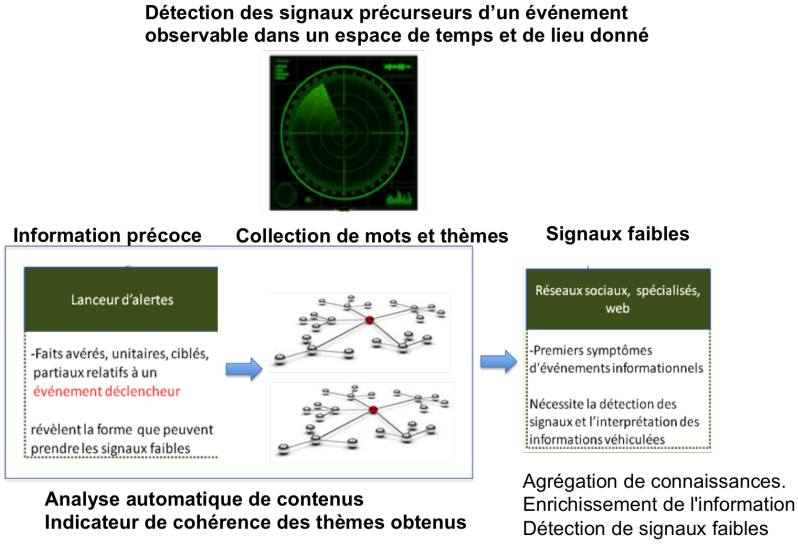


FIG. 2 – *Stratégie de détection des signaux faibles. Elle passe par l'analyse de l'information précoce apportée par un lanceur d'alerte et l'extraction des collections de mots associées aux thèmes découverts. Ces informations permettent ensuite de mieux cibler la phase de data mining pour la détection des signaux faibles.*

Cet article s'inscrit dans la première action. Afin de traiter de manière automatique les documents textuels révélés par le lanceur d'alerte et portant sur un ou plusieurs faits relatifs à l'événement déclencheur, nous développons des outils d'analyse afin de :

— regrouper les documents portant sur un même fait/thème ; le clustering *LDA* (Latent Dirichlet Allocation) permet, à partir de vecteurs descripteurs construits sur les documents, de relier ensemble avec un minimum d'*a priori* tous les documents relatifs à un même thème. Ces thèmes, que nous supposerons relatifs à l'événement déclencheur, sont découverts simultanément grâce au clustering multi-niveaux

— d'évaluer la qualité des partitions obtenues grâce à un indice de cohérence sur la collection de mots représentative de chaque thème obtenu. Deux algorithmes sont proposés. Le premier estime le nombre de thèmes le plus pertinent, et extrait ainsi sur ce niveau la collection de mots pour chacun des thèmes trouvés. Le second propose d'extraire la meilleure collection de mots pour chaque thème, celle-ci pouvant être potentiellement présente sur des niveaux différents. Ces mots et leurs attributs sont les indicateurs recherchés (lexique de descripteurs textuels). Ils seront utilisés par la suite lors des requêtes pour enrichir ce premier niveau d'information.

2 Clustering et text mining

La nécessité grandissante du traitement rapide de l'information conduit au développement de nombreux algorithmes utilisant diverses approches pour traiter les données. Des méthodes de traitement et d'extraction efficaces, comme *LDA* ou dérivées du "word embedding" sont régulièrement utilisées.
Le problème qui nous intéresse dans cette étude est celui de l'évaluation de l'efficacité du modèle *LDA*. Ce dernier catégorise les documents en un nombre de thèmes défini *a priori*. Afin d'améliorer la séparation en thèmes des documents, il est nécessaire (1) de faire varier ce paramètre, (2) d'estimer le niveau de partitionnement le plus pertinent, et (3), d'estimer dans l'arbre de profondeur la meilleures collection de mots représentative d'un thème. Pour cela nous nous appuyons sur une méthode récente de type "word embedding", *Word2Vec*. Celle-ci s'appuie sur une méthode d'apprentissage automatique issue du deep learning. Elle permet de représenter un mot par un vecteur dans un but d'analyse sémantique. Ainsi deux mots dans des contextes similaires ont des vecteurs proches. Cette approche s'avère donc complémentaire des méthodes s'appuyant sur le comptage de mots dans un document. Elle projette les mots dans un espace de vecteur en fonction du contexte local d'une phrase, au contraire du modèle *LDA* qui trie les mots en fonction de leurs probabilités dans les thèmes.
Nous commençons d'abord par décrire le fonctionnement de *LDA* ainsi que sa mise en oeuvre. Nous présentons ensuite l'approche *Word2Vec* et ce qu'elle apporte dans l'optimisation de *LDA*. Nous terminons par une présentation des deux algorithmes proposés et une discussion sur les résultats.

2.1 Quelques méthodes algébriques de représentation d'un document

Parmi l'ensemble des techniques de traitement des langues naturelles, l'analyse sémantique latente de Deerwester et al. (1990) (ou LSA) fait figure de pionnière. Cette technique décrit les relations entre les documents et les mots qu'ils contiennent. Une version probabiliste pLSA a servi d'inspiration pour le modèle LDA. pLSA de Hofmann (1999) intègre des techniques statistiques pour le traitement des mots où leurs composantes peuvent être considérées comme des représentations de «sujets». Chaque mot est ainsi généré à partir d'un seul sujet. Des variantes à LDA existent comme la méthode hiérarchique hLDA proposé par Blei et al. (2004).

2.2 Latent Dirichlet Allocation

Le modèle *LDA* est une méthode probabiliste générative de mots proposée par Blei et al. (2003) dont le but est découvrir les thèmes sous-jacents à un ensemble de documents. Chacun d'eux est modélisé par un mélange de thèmes générateur des mots du document. *LDA* est un modèle Bayésien à trois couches (cf. Figure 3). Elle utilise l'approche "Bag of Word" qui traite chaque document d du corpus D défini par $(\mathbf{w_1}, ..., \mathbf{w_D})$ comme un N-uplet de mots, $\mathbf{w_d} = (w_1, ..., w_N)$. A chaque mot $w_{(d,n)}$ est alors associé un thème représenté par la variable $z_{(d,n)}$. θ_d représente la distribution de thèmes du document d. Des hyperparamètres, α et η, définissent l'*a priori* sur θ et β où β_k décrit la distribution du thème k.

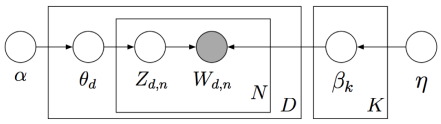


FIG. 3 – *Modèle graphique de* LDA *(Blei et al. (2003))*

LDA est un outil classiquement utilisé dans une grande variété de domaines : structuration automatique de corpus de documents, recommandation, génération de rapport de synthèse... Plusieurs recherches ont évalué l'efficacité de *LDA* dans des domaines où de grands volumes de données doivent être structurés thématiquement, notamment pour les réseaux sociaux, Hong et Davison (2010), le web profond, Noor et al. (2013) ou les encyclopédies numériques telle Wikipedia, Hoffman et al. (2010). Un grand nombre d'extensions de *LDA* ont également été proposé, par exemple, pour l'analyse sémantique latente probabiliste, PLSA, en traitement automatique des langues.
Dans la majorité des travaux, les utilisateurs utilisent *LDA* avec un nombre de thèmes *a priori*. Par exemple, Noor et al. (2013) utilise le dataset TEL-8, un ensemble de sources classé en 8 domaines ou bien Hoffman et al. (2010) qui cherche à structurer un ensemble d'articles de Wikipedia en 100 topics.

2.3 Mise en oeuvre. Application de *LDA* sur la base de connaissance Wikipedia

L'encyclopédie numérique Wikipedia français contient 3.9 millions de pages. Pour le choix du corpus et afin de construire une preuve de concept, nous avons choisi d'utiliser les documents de l"encyclopédie numérique Wikipedia français qui contient un nombre conséquent de documents (3.9 millions de pages). L'ensemble D de notre corpus de documents extrait de cette encyclopedie est constitué de 4 catégories génériques : "Economie", "Histoire", "Informatique", "Médecine" (cf. tableau 1). La structure des catégories dans Wikipedia est une arborescence particulière. Une feuille peut en effet se trouver sur plusieurs branches. De ce fait dans la version anglaise de Wikipedia, Bairi et al. (2015) explique qu'une catégorie principale couvre, si l'on explore ses sous-catégories jusqu'à une profondeur de 10, l'ensemble des documents présent dans l'encyclopédie.

	Nombre de pages	Echantillon de pages (0.6%)
Nombre de pages du Wikipedia français	3 987 661	/
Catégorie "Economie"	1 888 801	9 445
Catégorie "Histoire"	1 743 374	8 717
Catégorie "Informatique"	1 030 280	5 152
Catégorie "Médecine"	570 257	2 852

TAB. 1 – *Nombre de pages dans Wikipedia et dans les différentes catégories. Un échantillon de documents dans chaque catégorie est extrait et constitue notre corpus.*

Il existe de forts recouvrements entre les catégories, c'est pourquoi les pages spécifiques à un thème sont certainement peu nombreuses. Pour préserver une durée de calcul raisonnable, nous avons choisi de mener notre étude sur un sous échantillon du corpus, lequel se compose de 0.6% des documents de chaque catégorie. Cela représente un total de 26 000 documents. Chacune d'elle fait plus de 10 Ko. Le Tableau 2 montre le résultat de partitionnement par la méthode LDA lorsque le nombre de thèmes (ou clusters) k est fixé à 4.

	Thème 1	Thème 2	Thème 3	Thème 4
Mots	commune	film	saison	guerre
	ville	album	club	pays
	roi	premier	première	france
	nom	années	premier	général
	église	ans	tour	français
	...	...	...	...

TAB. 2 – *Liste des 5 premiers mots de chaque thème*

Pour rappel, LDA est une méthode de clustering (non supervisée) et ne permet donc pas d'associer une étiquette aux thèmes trouvés. De plus, il n'est pas possible de discerner la cohérence de chaque thème. Pour cela il est nécessaire de définir un indicateur, c'est l'objet de la prochaine section.

3 Word embedding

Le "word embedding" ou en français "plongement de mots" apporte une solution au problème de la dimensionnalité lié à la taille des dictionnaires. Cette approche permet d'une part de représenter les mots d'un dictionnaire par des vecteurs, et d'autre part, de prendre en compte la notion de contexte, facilitant l'analyse sémantique et syntaxique. Son implémentation s'appuie notamment sur les réseaux de neurones comme ceux présentés par Mikolov et al. (2013) qui permettent des estimations de probabilités significativement meilleures que les modèles n-grammes, (Mikolov et al. (2011), Bengio et al. (2003)). Actuellement, grâce à l'accroissement de la puissance de calcul (programmation GPU) et aux approches d'apprentissage profond, des

problématiques difficiles comme la Traduction, l'Analyse de sentiments ou la Reconnaissance vocale ont connu des avancés significatives.

3.1 Mise en oeuvre des indicateurs

Afin de définir notre indicateur, nous utilisons une méthode récente de type "word embedding" introduite par Mikolov et al. (2013), *Word2Vec*.
Dans Moody (2016), l'auteur décrit un algorithme hybride reposant sur les deux approches *LDA* et *Word2Vec*. L'algorithme porte le nom de *lda2vec*. Le vecteur du mots est estimé par *Word2Vec* en utilisant des informations locales représentées par les mots voisins, et des informations globales au corpus de documents apportées par *LDA*.

Notre approche, contrairement à celle de *lda2vec*, a comme objectif la recherche des clusters de mots ayant la plus forte similarité dans le contexte du corpus de documents. Elle s'appuie sur le fait que les mots sont représentés sous la forme de vecteurs caractéristiques des relations contextuelles qui les relient entre eux par l'intermédiaire de leur contexte (de voisinage). Il est alors possible de définir la valeur de similarité entre deux mots. Une valeur proche de 1 indique que les mots sont très proches l'un de l'autre (i.e. contexte semblable) et possède donc un lien sémantique fort. A l'inverse, 0 indique des mots peu employés dans des contextes semblables. Nous utilisons cette valeur de similarité pour construire un indicateur de cohérence. Celui-ci se définit comme la somme des valeurs de similarité de toutes les combinaisons de mots deux à deux dans chacun des thèmes. Il est donné par l'équation (1) et permet ainsi d'analyser la cohérence d'un thème. E est l'ensemble des 100 mots supports du thème, P l'ensemble des k-combinaisons de E et $w2vSim$ la mesure de similarité définie dans *Word2Vec* par Mikolov et al. (2013). Plus la valeur est grande et plus le thème contient des mots régulièrement employés ensembles.

$$ind = \sum w2vSim(P_{k=2}(E)) \tag{1}$$

L'étape suivante consiste à appliquer l'algorithme *LDA* en faisant évoluer le nombre de clusters/thèmes. Nous obtenons ainsi plusieurs partitions calculées sur un nombre de thèmes différents, et qu'il est possible de représenter sous la forme d'une arborescence. Il est à noter que *LDA* ordonne les thèmes découverts lors des différentes itérations dans un ordre aléatoire, une étape supplémentaire est donc nécessaire pour construire cette arborescence (voir ci-après). Les résultats sont présentés sur le tableau 3. k représente le nombre de clusters donné en entrée de *LDA*.

	Thème 1	Thème 2	Thème 3	Thème 4	Thème 5	Thème 6
2 thèmes (k=2)	1367	2469				
3 thèmes (k=3)	1337	1867	2487			
4 thèmes (k=4)	1480	2052	2356	1948		
5 thèmes (k=5)	2104	1633	3284	1921	1416	
6 thèmes (k=6)	2070	3181	2013	1382	2051	1820

TAB. 3 – *Evolution de l'indicateur en fonction du nombre de thèmes k donné en entrée de* LDA

Afin d'évaluer les partitions obtenues, nous proposons deux algorithmes s'appuyant sur l'indicateur précédent. Le premier (décrit en Section 3.2) estime le nombre de thèmes k (i.e. le niveau de l'arborescence) le plus pertinent, et extrait ainsi sur ce niveau, la collection de mots pour chacun des thèmes trouvés. Le second (décrit en Section 3.3) propose d'extraire les meilleurs collections de mots potentiellement présentes sur des niveaux différents.
Afin de construire l'arborescence, il est nécessaire d'évaluer le lien de ressemblance entre des thèmes de niveaux différents. Celui-ci se calcule au moyen d'un indicateur de ressemblance (2). Pour l'ensemble, C, des mots communs w présents à la fois dans un thème de niveau n, T_n, et un thème de niveau $n+1$, T_{n+1}, nous calculons la somme normalisée du produit des probabilités $p_{.,l}$ associés à ces mots commun dans les deux thèmes (cf. Figure 4).

$$R(T_{n+1}, T_n) = \frac{\sum_{l \in C} p_{n,l} \cdot p_{n+1,l}}{\sum_{l \in T_{n+1}} p_{n+1,l}^2} \tag{2}$$

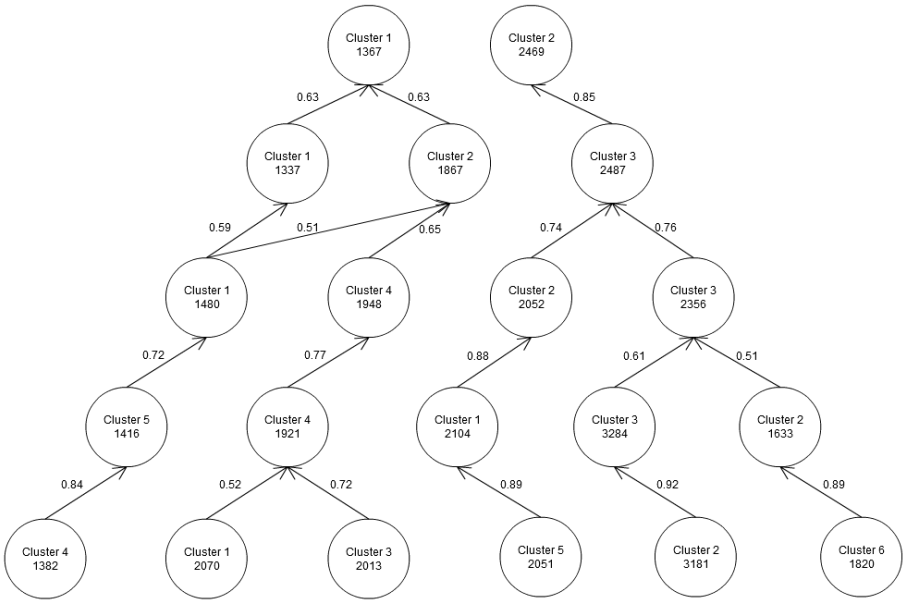


FIG. 4 – *Partitions et arborescence obtenus sur K niveaux (K fixé à 6 dans cet exemple) avec* LDA. *Représentation des liens entre les thèmes au moyen de l'indicateur de ressemblance. Ne sont gardés que les liens dont la valeur est supérieure à 10.*

3.2 Recherche du k le plus pertinent

L'algorithme 1 consiste en la recherche du niveau de l'arborescence donnant les thèmes les plus cohérents au sens de l'indicateur Eq. (1). Sur chaque niveau, nous calculons la valeur minimale de cet indicateur sur l'ensemble des clusters présents sur le niveau. Le niveau retenu (et donc le nombre de clusters pertinent au sens du critère) correspond à celui dont la valeur

mininale est la plus grande. La figure 5 montre les différents clusters. Le cluster dont la cohérence est la plus grande parmi les clusters les moins cohérents de chaque niveau est le thème 1 du niveau 4 avec la valeur 1480 (cf. Figure 5).

3.3 Recherche des clusters les plus pertinents sur l'ensemble de l'arborescence *LDA*

Il est possible d'extraire sur toute l'arborescence les clusters/thèmes les plus pertinents au sens du critère de cohérence, $ind(T)$, et des relations de ressemblance, $R(T_n, T_{n+1})$, entre un cluster de niveau n et de niveau $n + 1$. Nous proposons pour cela une méthode de parcours des thèmes dans l'arborescence de manière ordonnée selon le critère ind(T), où chaque thème nouvellement rencontré est retenu comme pertinent et entraine le retrait dans l'arborescence de tous ses thèmes parents ou fils. Les liens de parenté entre les thèmes (décris par R) ne sont considérés qu'au delà d'un seuil fixé arbitrairement à 0,5. L'algorithme 2 formalise ce parcours.

3.4 Interprétation des résultats

Nous avons partitionné les documents en un nombre de thèmes. La recherche des clusters/thèmes les plus pertinents a permis de mettre en évidence la nécessité de faire varier la valeur de k passée en entrée de la méthode LDA. En effet, selon la valeur de k, nous obtenons des clusters de mots avec de fortes valeurs de cohérence et d'autres moins. Le cluster 3 dans la partition à 5 thèmes contient les mots [saison, club, france, match, championnat]. On remarque que ce sont essentiellement des mots en relation avec le sport. Au contraire du cluster 6 dans la partition à 6 thèmes qui contient des mots ayant peu de liens entre eux comme [autres, cas, exemple, système, certains, plusieurs]. On remarque que l'on ne peut pas définir d'étiquette à ce regroupement de mots. En fixant une valeur de seuil (par exemple 2000), nous pouvons identifiter 2 clusters de rejet contenant des mots ayant peu de similarité. Ces clusters de rejet contiennent des groupes de mots qui ont de fortes valeurs de cohérences [commune, ville, région, département, population], mais sont finalement peu nombreux dans le cluster. Un repartitionnement de ces clusters de rejet permettrait de mettre en evidence ces relations faibles.

4 Conclusion

Dans cet article, nous avons proposé une approche pour la recherche de thèmes/faits communs au sein d'un corpus de documents. La combinaison LDA / word2vec telle que nous avons proposé de la mettre en oeuvre permet de s'affranchir du paramètre k (nombre de clusters) pour le partitionnement. Deux directions ont été explorées : 1) un premier algorithme (c.f. Section 3.2) visant à rechercher le nombre de thèmes (paramètre k) entraînant un partitionnement par LDA le plus cohérent possible ; 2) un algorithme (c.f. Section 3.3) qui, de manière plus avancée, combine les meilleurs thèmes renvoyés par LDA sur l'ensemble des partitionnements (ou valeurs de k) testées.
La valeur de seuil fixée pour le parcours et l'élagage de l'arborescence a été arbitrairement fixée à 10. Les valeurs supérieures matérialisent une relation forte entre les thèmes, alors que

les valeurs inférieures peuvent être assimilées à des relations moins évidentes, mais pourtant bien existantes. Actuellement considérés comme des clusters de rejet, ces thèmes et relations aussi infimes soient-elles peuvent éventuellement matérialiser des signaux faibles. Dans le contexte de notre étude sur la détection des signaux faibles et de lancements d'alertes, nous pensons que ces signaux / relations faibles méritent d'être étudiés.
L'information portée par ces derniers devra être corrélé à un contexte informationnel plus large au moyen de phase d'exploration sur les réseaux. Ceci dans un objectif de détection de signaux précurseurs dont la présence attenante dans un espace de temps et de lieux donnés anticipe l'avènement d'un fait observable.

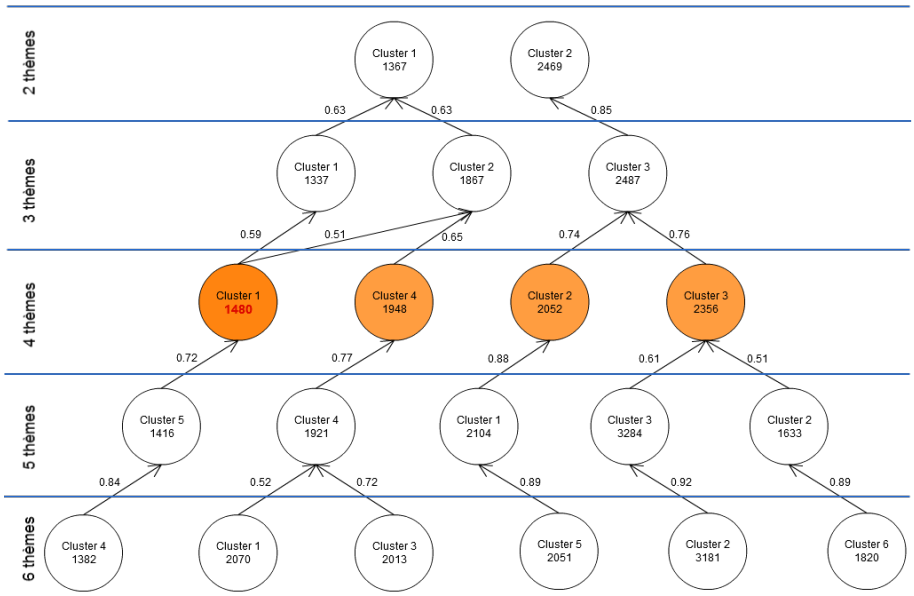


FIG. 5 – *Application de l'algorithme* (1) *sur l'arborescence* LDA *obtenue*

Data : P = Liste des nombres de clusters demandés : {2...K}
Result : bestk = identifiant du k niveau
bestk $\leftarrow$ 0;
bestScorek $\leftarrow$ `Min(LDA(`bestk`))`;
for $k \in P$ **do**
  if `Min(LDA(`k`))` $>$ bestScorek **then**
    bestk $\leftarrow$ k;
    bestScorek $\leftarrow$ `Min(LDA(`k`))`;
  end
end
return bestk

Algorithme 1 : Récupération de l'identifiant du niveau k optimal

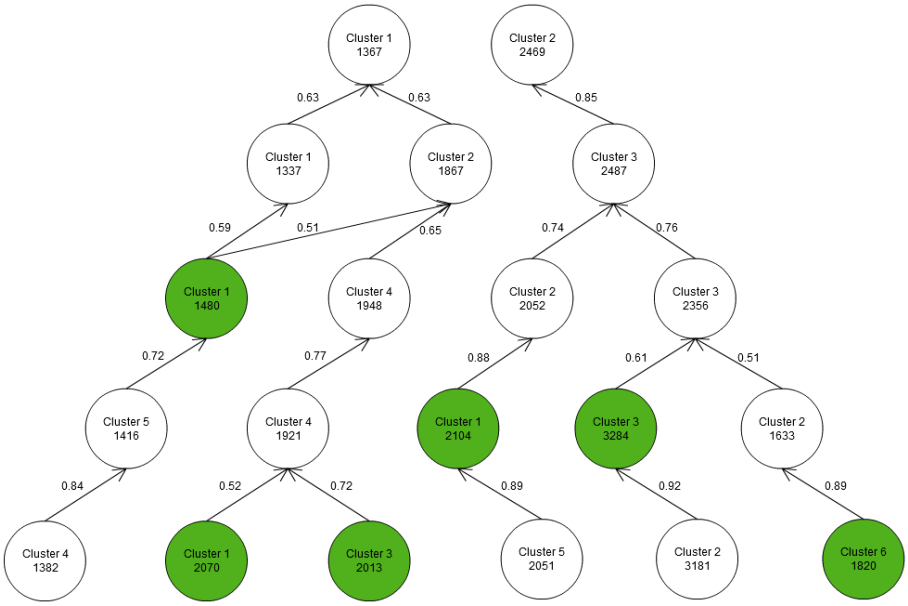


FIG. 6 – *Application de l'algorithme* (2) *sur l'arborescence* LDA *obtenue*

Data : T = Liste des thèmes de l'arborescence *LDA* triés par valeur de cohérence
Result : themesRetenus = Liste des identifiants des thèmes pertinents
themesRetenus $\leftarrow$ {};
while `Taille(`T`)` > 0 **do**
  meilleurCluster $\leftarrow$ `Max(`T`)`;
  themesRetenus $\longleftarrow$ themesRetenus + {meilleurCluster };
  for $t \in$ `Parents(`meilleurCluster`)` **do**
    $T \longleftarrow T - t$;
  end
  for $t \in$ `Fils(`meilleurCluster`)` **do**
    $T \longleftarrow T - t$;
  end
end
return themesRetenus

Algorithme 2 : Récupération des thèmes pertinents dans l'arborescence *LDA*

Références

Bairi, R. B., M. Carman, et G. Ramakrishnan (2015). On the Evolution of Wikipedia : Dynamics of Categories and Articles. *2015 ICWSM Workshop*, 1–8.

Bengio, Y., R. Ducharme, P. Vincent, et C. Janvin (2003). A Neural Probabilistic Language Model. *The Journal of Machine Learning Research 3*, 1137–1155.

Blei, D., T. Griffiths, M. Jordan, et J. Tenenbaum (2004). Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems 16*.

Blei, D. M., A. Y. Ng, et M. I. Jordan (2003). Latent Dirichlet Allocation. *Journal of machine learning research : JMLR 3*, 993–1022.

Deerwester, S., S. T. Dumais, et R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American society for information science 41*(6), 391–407.

Hoffman, M. D., D. M. Blei, et F. Bach (2010). Online Learning for Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems 23*, 1–9.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*, 50–57.

Hong, L. et B. D. Davison (2010). Empirical study of topic modeling in twitter. *Proceedings of the First Workshop on Social . . .*, 80–88.

Mikolov, T., G. Corrado, K. Chen, et J. Dean (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, 1–12.

Mikolov, T., A. Deoras, S. Kombrink, L. Burget, et J. H. Černocký (2011). Empirical evaluation and combination of advanced language modeling techniques. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 605–608.

Moody, C. (2016). Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec.

Noor, U., A. Daud, et A. Manzoor (2013). Latent Dirichlet Allocation based Semantic Clustering of Heterogeneous Deep Web Sources.

Summary

This paper is related to a wide project aiming at discoving from different streams of information (i.e. daily publication from the Internet), weak signals possibly sent by whistleblowers. The current study presented in this paper tackles the particular problem of clustering topics at multi-levels from multiple documents, and then extracting meaningful descriptors, such as weighted lists of words. In this context, we present a novel idea combining LDA (in charge clustering) and Word2vec (providing a consistency metric regarding the partitioned topics) as potential method for limiting the $apriori$ number of cluster k usually needed in classical partitioning approaches. We proposed 2 implementations of this idea, respectively able to: (1) finding the optimal k for LDA ; (2) gathering the optimal clusters from different levels of clustering.

Erreurs OCR et biais d'indexation : impact sur les usages

Guillaume Chiron*,***, Jean-Philippe Moreux*,****
Antoine Doucet**,♯, Mickaël Coustaty**,♯♯, Muriel Visani**,♯♯♯

*Bibliothèque nationale de France, Service numération, Paris
**L3i, Université de la Rochelle, Avenue Michel Crépeau, 17042 La Rochelle
***guillaume.chiron@bnf.fr, ****jean-philippe.moreux@bnf.fr
♯antoine.doucet@univ-lr.fr, ♯♯mickael.coustaty@univ-lr.fr, ♯♯♯muriel.visani@univ-lr.fr

Résumé. Les méthodes d'analyse classiquement appliquées dans le contexte du *Big Data*, provoquent souvent un phénomène de « boîte noire » où la qualité de numérisation des documents peut être un paramètre négligé. En dépit des bonnes pratiques en vigueur inhérentes au métier de *data-journalist*, se pose la problématique des biais statistiques induits par ce manque de transparence sur la fiabilité des sources. S'inscrivant dans le cadre du projet AméliOCR, cet article vise à estimer ces potentiels biais sur l'indexation et la recherche. Cette étude s'appuie sur un corpus de documents OCéRisés associés à leur vérité terrain, ainsi que sur des historiques de recherche sur Gallica.

1 Introduction

L'amélioration des technologies de numérisation – toutes sources confondues (p. ex. documents papiers, archives audio/vidéo) – associée à des méthodes de traitement toujours plus performantes (p. ex. OCR/reconnaissance de textes/images, transcription automatique) génère une quantité croissante d'informations. Pour les *data-journalists*, cela constitue un terrain de jeu au potentiel sans précédent, mais dans lequel il est recommandé de s'aventurer avec précaution. Les méthodes d'analyses (p. ex. filtrage, croisement de données) classiquement appliquées dans le contexte du Big Data, négligent souvent l'aspect qualitatif des documents numériques exploités pour ne favoriser qu'une approche quantitative imparfaite.

Dans cet article, nous nous intéressons au cas particulier des documents textuels OCéRisés. Ce travail s'inscrit dans la cadre du projet AméliOCR [1] lancé en 2016, dont l'objectif est d'améliorer la qualité du texte dans les documents historiques numérisés au sein de la bibliothèque numérique Gallica [2]. Nous portons une attention particulière à détecter et corriger les erreurs d'OCR qui affectent les termes les plus recherchés dans Gallica. L'enjeu est de taille, sachant l'impact que des mauvais résultats de recherche peuvent avoir sur des analyses automatisées (Traub et al., 2015). À titre d'exemple, on citera l'utilisation de Gallica comme source d'attestation de lexique [3] et ce cas emblématique de biais : une recherche sur « gadget »

1. Fruit d'une collaboration entre la Bibliothèque nationale de France et le laboratoire L3i.
2. Bibliothèque numérique de la BnF en libre accès : http ://gallica.bnf.fr
3. « Alain Rey et Gallica : une grande histoire de mots », http ://gallica.bnf.fr/blog/20102016/alain-rey-et-gallica-une-grande-histoire-de-mots

(à l'étymologie discutée) dans la presse du XIXe renvoie de nombreuses occurrences qui sont en fait des transcriptions OCR erronées de « budget » !

On retrouve dans la littérature un certain nombre de travaux visant à améliorer *a posterori* des résultats d'OCR. Certains s'appuient essentiellement sur des modèles de langages tels que (Bassil et Alwani, 2012) via *Google Suggest*, d'autres utilisent des modèles d'erreurs (Brill et Moore, 2000) voire éventuellement avec un retour à l'image (Lee et Smith, 2012). Ces approches se heurtent généralement à des limites statistiques (Smith, 2011), et c'est d'ailleurs pour cela que bon nombre d'initiatives de correction assistées par l'homme ont été proposées (Taghva et Stofsky, 2001). Néanmoins le problème, notamment pour les documents anciens qui sont particulièrement difficiles à OCéRiser, reste entier.

La première phase du projet AméliOCR nous amène à proposer deux contributions : 1) la constitution d'un corpus pour l'analyse des erreurs d'OCR, qui sera prochainement rendu public ; 2) une méthode d'alignement entre les textes OCéRisés et leur vérité terrain [4].

2 Constitution d'un corpus OCR / VT

1) Compilation des documents – Nous avons rassemblé un corpus de documents anciens en français qui est à notre connaissance le plus conséquent dans son genre. Comme le montre le tableau 1, il regroupe des documents OCéRisés de natures différentes (p. ex. journaux, monographies) dont certains proviennent de projets de recherche européens antérieurs (IMPACT, Europeana Newspapers) et d'autres de projets ou programmes de numérisation de la BnF. La richesse de ce corpus est qu'il dispose pour chaque document OCéRisé (OCR) d'une vérité terrain (VT) sur le texte. La majorité des documents (OCR + VT) sont disponibles en libre accès, mais reposent sur une variété de formats et de versions utilisés dans le domaine (p. ex. ALTO, PAGE, EPUB, texte brut) ainsi qu'un certain nombre de spécificités propres à chacun (p. ex. métadonnées, encodages, ordre de lecture renseigné ou non). Afin de rendre ce corpus accessible d'une part, et permettre l'agrégation de ces données au sein d'un corpus homogène, un important travail d'ingénierie a été réalisé.

Source	**Nature**	**Dates**	**Symboles alignés**
Europeana News. (52 pages)	périodiques	1814 - 1944	1 066 994 (92%)
IMPACT (1004 pages)	monographies	1821 - 1864	1 190 331 (98%)
VT BnF (6656 pages)	mixte	1820 - 1943	8 861 428 (98%)
Marché de masse (151 pages)	mixte	1654 - 2000	270 471 (95%)
Presse autre (32 pages)	périodiques	1897 - 1934	650 720 (90%)
Monog autre (70 livres)	monographies	1610 - 1926	16 518 313 (99%)
	TOTAL	1610 - 2000	27 597 957 (98,7%)

TAB. 1 – *Constitution du corpus FR pour le projet AméliOCR*

2) Alignement OCR / VT – Pour pouvoir identifier les erreurs-types d'OCR et analyser leur fréquence, il est nécessaire d'aligner au symbole près les deux versions. Les outils d'alignement traditionnellement utilisés par la communauté tels que ISRI (Rice et Nartker, 1996) – voire d'autres extensions [5] – n'ont pas été conçus pour gérer des séquences à l'échelle d'un

4. Texte transcrit à la main à partir des images.
5. `https://github.com/kba/awesome-ocr`

livre. Des approches récentes telles que (Yalniz et Manmatha, 2011; Al Azawi et al., 2013), plus performantes, offrent une solution à ce problème. Celles-ci fonctionnent de manière récursive via l'identification de sous-chaînes similaires de plus en plus petites entre l'OCR et la VT. Le positionnement d'ancres à différentes échelles permet ainsi un appariement allant jusqu'au niveau du caractère. Ce mécanisme d'ancrage strict peut poser problème lorsque l'OCR est trop dégradé, faute de pouvoir identifier suffisamment de sous-chaînes similaires. C'est pourquoi nous avons développé une approche d'alignement à base d'ancrages flous où des sous-chaînes légèrement différentes peuvent servir de repère pour l'alignement. Les séquences non-similaires (par opposition aux séquences similaires) entre les repères sont ensuite alignées à l'aide d'une méthode originalement utilisée pour l'alignement de paires d'ADN (Smith et Waterman, 1981). En dépit d'un temps de calcul plus important, cette nouvelle méthode conduit à des résultats d'alignement satisfaisant sur les parties bruitées de l'OCR. Au total, le corpus a été aligné à plus de 98%, ce qui représente près de 27,5 millions de symboles appariés. Cela permet entre autres de retrouver les erreurs réalisées par les logiciels de reconnaissance de caractères du marché d'une part, et les termes originaux impliquées d'autre part (cf. fig 1). Par exemple, nous avons pu constater que les mesures de la qualité renseignées par certains moteurs OCR sont décorrélées des erreurs réellement avérées. Ces métriques (p. ex. « Word Confidence » renvoyé par l'OCR FineReader) sont donc peu fiables dans un contexte d'exploitation à grande échelle via une approche automatique.

IMG	rappelle aux jeune gens qui ont satisfait aux examens prescrits par										
OCR	rappelle	aux	jeune#	gens	qui	on*	#####fait	aux	examens	prcscrits	par
VT	rappelle	aux	jeunes	gens	qui	ont	satisfait	aux	examens	prescrits	par
WC	0.70	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.99	1.00	0.99

FIG. 1 – *Alignement entre l'OCR et la VT, où le "#" symbolise les symboles manquants (ou illisibles). Word Confidence (WC) est une métrique exprimant le taux qualité estimé au mot.*

3 Croisement entre les erreurs et les recherches dans Gallica

L'analyse porte sur les deux points suivants : 1) la nature et la fréquence des erreurs d'OCR ; 2) le croisement de ces erreurs avec les termes les plus fréquemment recherchés sur Gallica :

1) Fréquence des erreurs – La figure 2 donne un aperçu des erreurs d'OCR les plus fréquentes. Sur les 5 millions de mots constituant le corpus, on compte plus de 100k mots concernés par des erreurs d'OCR (hors erreurs de ponctuation), ce qui représente 2 % des mots corpus. Parmi ces erreurs, on estime que 1) 15 % des mots mals OCéRisés sont des noms propres (repérés ici par une majuscule non précédée d'un point), et que 2) près de la moitié des erreurs concerne des termes non présents dans un dictionnaire classique (i.e. dictionnaire OpenOfficeFr).

2) Croisement avec les recherches Gallica – Nous exploitons pour cela la liste des 26 000 termes les plus fréquemment requêtés sur Gallica sur une période de 4 mois (de décembre 2015 à mars 2016). On note qu'un nombre important de noms propres – soit environ 79% sur les 500 premières requêtes – sont la cible de recherches. Globalement, nous observons que 21% des mots (ou occurrences) de la vérité terrain se retrouvent dans les logs de recherche. Ce chiffre important s'explique par les termes communément utilisés (p. ex. « sont », « pour »). Parmi les

1 symbole		
OCR	VT	Freq.
f	s	5069
l	t	1915
c	e	1633
o	e	1344
i	l	1252
	s	1074
t	l	1066
u	n	1007

2 symboles		
OCR	VT	Freq.
fi	st	286
li	h	264
ff	ss	246
11	ll	226
fl	st	223
in	m	186
ni	m	178
ii	n	169

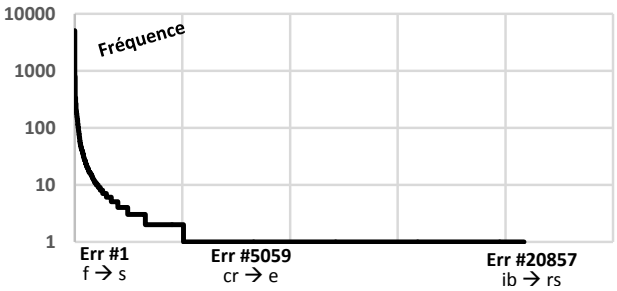


FIG. 2 – *Fréquences des 20857 erreurs d'OCR constatées sur le corpus, avec un tableau détaillé montrant les 8 premières touchant 1 et 2 symboles.*

termes présents à la fois dans les logs de Gallica et dans la VT (au nombre de 3492), 1% sont victimes d'erreurs d'OCR. Voici quelques exemples de termes recherchés sur Gallica les plus fréquemment mal OCéRisés, ainsi que le nombre d'occurrences de ces erreurs :
- *TERMES FRÉQUENTS = [sont*420, sous*199, pour*194, dans*176, cette*150, ...]*
- *ENTITÉS NOMMÉES = [France*35, Egypte*27, Rome*17, Russie*12, Edouard*11, ...]*

4 Conclusion

Cet article présente les prémisses du travail effectué dans le cadre du projet AméliOCR qui vise à proposer une approche de correction automatique des erreurs d'OCR sur les documents numérisés. Les enjeux sont la réduction des biais d'indexation et de recherche, lesquels peuvent altérer les résultats d'analyses menées en aval par les utilisateurs de portails documentaires ou de corpus numériques.

Ainsi, la première phase du projet a donné lieu à deux contributions : 1) la constitution d'un corpus pour l'analyse des erreurs d'OCR ; 2) une approche d'alignement entre les textes OCéRisés et leur vérité terrain. Bien que les chiffres – issus de calculs automatisés – soient à considérer comme des estimations, ce travail vise à alerter les professionnels de l'information des possibles biais rencontrés lors d'analyses massives et automatisées de corpus numérisés, en dépit des bonnes pratiques en vigueur dans leur métier.

Références

Al Azawi, M., M. Liwicki, et T. M. Breuel (2013). Wfst-based ground truth alignment for difficult historical documents with text modification and layout variations. In *IS&T/SPIE Electronic Imaging*, pp. 865818–865818. International Society for Optics and Photonics.

Bassil, Y. et M. Alwani (2012). Ocr post-processing error correction algorithm using google's online spelling suggestion. *Journal of Emerging Trends in Comp. and Info. Sciences 3*.

Brill, E. et R. C. Moore (2000). An improved error model for noisy channel spelling correction. In *Proceedings of Annual Meeting on Association for Comp. Linguistics*, pp. 286–293.

Lee, D.-S. et R. Smith (2012). Improving book ocr by adaptive language and image models. In *Document Analysis Systems, 10th IAPR International Workshop on*, pp. 115–119. IEEE.

Rice, S. V. et T. A. Nartker (1996). The isri analytic tools for ocr evaluation. *UNLV/Information Science Research Institute, TR-96-02*.

Smith, R. (2011). Limits on the application of frequency-based language models to ocr. In *2011 International Conference on Document Analysis and Recognition*, pp. 538–542. IEEE.

Smith, T. F. et M. S. Waterman (1981). Identification of common molecular subsequences. *Journal of molecular biology 147*(1), 195–197.

Taghva, K. et E. Stofsky (2001). Ocrspell : an interactive spelling correction system for ocr errors in text. *International Journal on Document Analysis and Recognition 3*(3), 125–137.

Traub, M. C. et al. (2015). Impact analysis of ocr quality on research tasks in digital archives. In *International Conf. on Theory and Practice of Digital Libraries*, pp. 252–263. Springer.

Yalniz, I. Z. et R. Manmatha (2011). A fast alignment scheme for automatic ocr evaluation of books. In *2011 International Conference on Document Analysis and Recognition*, pp. 754–758. IEEE.

Summary

The processing methods conventionally applied in Big Data often cause a "black box" effect into which the quality of digitized documents is often neglected. Despite the good practices of data-journalists, arises the problem of statistical biases induced by this lack of transparency on the reliability of the sources. As part of the AméliOCR project, this study aims to estimate these potential biases on indexing and searching. This work is based on a corpora of OCR-ized documents associated with their ground truth, as well as historical search logs gathered from Gallica.

Index

Aubrun, Frédéric, 1

Bouju, Alain, 57

Cagé, Julia, 25
Chiron, Guillaume, 57, 69
Coustaty, Mickael, 69

Del Vecchio, Myrian, 1
Doucet, Antoine, 69

Ghoniem, Mohammad, 41
Grabar, Natalia, 45

Hervé, Nicolas, 25

Kurpiel, Solange, 1

Labbé, Cyril, 13

Médoc, Nicolas, 41

Maitre, Julien, 57
Mazoyer, Béatrice, 37
Menard, Michel, 57
Moreux, Jean-Philippe, 69

Nadif, Mohamed, 41

Otavio, Luis, 1

Portet, François, 13

Richey, Mason, 45

Soulages, Jean-Claude, 1

Turenne, Nicolas, 37

Velcin, Julien, 1
Viaud, Marie-Luce, 25, 37
Visani, Muriel, 69
Vizzini, Jérémy, 13