Le traitement automatique des langues face aux données textuelles volumineuses et potentiellement dégradées : qu'est-ce que cela change ?

Pascale Sébillot

▶ **To cite this version:**

Pascale Sébillot. Le traitement automatique des langues face aux données textuelles volumineuses et potentiellement dégradées : qu'est-ce que cela change ?. Lisette Calderan; Pascale Laurent; Hélène Lowinger; Jacques Millet. Big data : nouvelles partitions de l'information. Actes du séminaire IST INRIA„ octobre 2014, De Boeck, pp.43-60, 2015, Information et stratégie, 978-2804189150. hal-01056396

HAL Id: hal-01056396
https://hal.science/hal-01056396

Submitted on 21 Aug 2015

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.

Le traitement automatique des langues face aux données textuelles volumineuses et potentiellement dégradées : qu'est-ce que cela change ?

Pascale Sébillot

1. Introduction

Prétendre que le phénomène récent du *Big Data* a bouleversé théoriquement et méthodologiquement le traitement automatique des langues (TAL) serait inexact : le TAL a connu sa dernière véritable révolution à la fin des années 80 - début des années 90. C'est en effet à cette période que s'est opéré un changement de paradigme, un passage du rationalisme vers l'empirisme, avec le « remplacement » des approches symboliques, à base de règles, fondées sur l'expertise humaine par des approches empiriques fondées sur les données, où la connaissance est extraite par des techniques d'apprentissage automatique, en particulier statistique. Permis par des capacités de stockage et de traitement accrues des ordinateurs et la disponibilité de volumes conséquents de textes au format numérique, ce bouleversement, même s'il s'est étendu sur plusieurs années, a été en fait assez profond : passage d'un TAL très linguistique où l'on cherchait à comprendre – expliquer les jugements de grammaticalité dont sont capables les locuteurs natifs d'une langue, construire et manipuler des représentations assez élaborées du sens… – à un TAL « très statistique » où l'on fait émerger des connaissances grâce à l'observation à grande échelle, au comptage…, et où l'on extrait des représentations de « sens utile » pour des traitements applicatifs.

Affirmer que les volumes actuels de données à l'échelle du *Big Data* n'ont eu et n'ont aucun impact sur le TAL serait cependant également faux. Les données textuelles à traiter se déclinent à l'aune des 3 V (variété, volume, vélocité). Elles consistent bien sûr en des documents écrits, pages Web, *emails* et autres textes « traditionnels », mais également en contenus de blogs, de réseaux sociaux, en *sms*, en documents audio transcrits automatiquement, ce qui correspond donc à des types et des qualités de langue très divers. Pour ne citer que quelques chiffres donnant tant une idée des volumes que de la vitesse d'évolution de ceux-ci, nous pouvons par exemple nous référer à ceux de *Go-globe.com* de juin 2011, cités à *davidfayon.fr/2011/croissance-du-web-une-minute*, qui, quoique un peu anciens, sont déjà extrêmement parlants : création par minute de 60 blogs, de 98000 messages sur Twitter, de 510000 commentaires sur Facebook, de 168 millions d'*emails* ou de 600 vidéos sur YouTube. À l'heure actuelle, ces nombres sont encore plus impressionnants. Ces volumes énormes de données textuelles ont accru le phénomène décrit précédemment de passage du TAL à l'empirisme, accompagné du renforcement de certains champs du domaine – parfois avec un regard nouveau – et de l'émergence d'applications nouvelles.

Dans la suite de ce chapitre, nous revenons en détail sur l'impact de ce déluge de données sur le TAL en débutant par un rappel des spécificités des données textuelles au sein de ce monde du *Big Data* dans lequel les données volumineuses auxquelles il est fait référence sont fréquemment (semi-)structurées ; ceci nous permet de mieux comprendre l'intérêt mais aussi la difficulté d'accéder au contenu sémantique de ces données particulières. Nous nous penchons ensuite sur la façon dont les chercheurs en TAL représentent et exploitent ces données massives pour en faire émerger la connaissance utile pour l'objectif visé. Nous présentons ensuite successivement d'une part des applications qui tentent de trouver des solutions pour faire face au déluge de données disponibles,

d'autre part certaines qui, elles, cherchent à tirer profit de cette masse d'informations et à exploiter sa redondance. Nous concluons en rappelant les grandes lignes de l'évolution du TAL.

2. Les données textuelles : types et caractéristiques

Dans [10], Halper *et al*. indiquent qu'actuellement environ 80% des données à disposition des entreprises sont des données textuelles non structurées, insistant sur la nécessité pour elles de savoir les exploiter. Le terme « données textuelles » recouvre cependant, tant dans le monde de l'entreprise que dans un cadre plus général, des réalités très diverses. L'objectif de cette section est donc, dans un premier temps, de revenir sur les 3 V du *Big Data* appliqué aux données textuelles et, ayant déjà abordé en introduction les idées de volumes et vélocité, de s'attarder plus particulièrement sur la notion de variété des données impliquées et de variabilité de leur qualité. Ceci nous conduit naturellement vers la mise au jour de leurs caractéristiques qui complexifient l'accès à leur contenu sémantique.

2.1 Types de données textuelles

Les données textuelles dont il est question dans ce chapitre sont donc multiples en termes de types et de qualité (un autre critère de variation est le multilinguisme que nous ne traitons pas spécifiquement).

Un premier regroupement des données disponibles concerne celles qui sont directement produites sous une forme textuelle. C'est le cas des textes au format numérique, des pages Web, des *emails*, des blogs, de certains contenus de réseaux sociaux ainsi que des *sms*. Si l'on peut considérer que l'accès au contenu textuel des premiers d'entre eux, respectant à peu près les règles habituelles de syntaxe et d'orthographe, est *a priori* assez aisé, il faut noter que les pages Web nécessitent souvent la mise en place d'un processus de nettoyage qui peut aboutir à une perte de certains éléments de structuration. Au fur et à mesure que l'on progresse dans la liste, la qualité rédactionnelle se modifie (orthographe potentiellement déficiente, syntaxe très relâchée…). Les *sms* sont un exemple extrême de cette déformation, mêlant une écriture standard à des formes phonétisées, des émoticônes, et un emploi de règles plus ou moins implicites respectées par les adeptes de ce mode d'échange, ce qui rend leur interprétation parfois très difficile même pour un humain [6]. Ce langage *sms* fait l'objet de diverses études comme au sein du projet international *sms4science* visant à constituer des *corpora* suffisamment vastes pour en étudier les spécificités.

Le second groupe concerne les données textuelles issues d'un média différent, dont la transformation en texte nécessite l'utilisation d'un système de reconnaissance, inévitablement générateur d'erreurs. C'est par exemple le cas des images de textes (dactylographiés, manuscrits…) transformées en texte grâce à un logiciel de reconnaissance optique de caractères. Ces logiciels procèdent habituellement en deux étapes : une segmentation de l'image en caractères individuels, suivie d'une reconnaissance de ceux-ci à l'aide d'un classifieur souvent fondé sur une approche neuronale. Les données textuelles peuvent aussi provenir de l'application d'un système de reconnaissance automatique de la parole à des documents audio ou vidéos. Dans le cadre de la modélisation statistique de la parole, la tâche de transcription équivaut à rechercher parmi l'ensemble des séquences de mots possibles, défini à partir d'un vocabulaire fixé, la séquence la plus probable étant donnée une séquence de caractéristiques acoustiques observées à partir du signal d'entrée. Elle repose sur quatre modules : un module de caractérisation du signal transformant le signal audio en une séquence de caractéristiques numériques, un vocabulaire définissant l'ensemble des mots pouvant être reconnus, un modèle acoustique calculant la vraisemblance du signal représenté par la séquence de caractéristiques sachant une séquence de mots, et un modèle de langue permettant de calculer la probabilité *a priori* d'une séquence de mots.

Les transcriptions produites se distinguent toutefois d'un texte classique selon plusieurs aspects : elles ne contiennent pas de ponctuation, sont, dans la plupart des systèmes, en minuscules, ne sont pas structurées en phrases mais en groupes de souffle correspondant à peu près à la parole prononcée par un locuteur entre deux respirations, et incluent un nombre potentiellement élevé de mots mal transcrits. La qualité d'une transcription est souvent mesurée grâce au taux d'erreurs de mots, distance minimale d'édition entre la transcription et une transcription de référence, rapportée au nombre de mots de la référence. Outre la séquence de mots la plus probable étant donné le signal d'entrée, un système de reconnaissance peut produire d'autres sorties, dont les mesures de confiance qui sont des scores, compris entre 0 et 1, associés à chaque mot généré et d'autant plus élevés que le système estime fiable la transcription de ce mot.

Les données textuelles auxquelles les méthodes et outils du TAL ont à se confronter actuellement sont donc variées, de qualité potentiellement très dégradée et très volumineuses. Pour donner à nouveau une idée de volume et de croissance récente, nous pouvons citer les 500 millions de *tweets* journaliers mentionnés par *www.blogdumoderateur.com/chiffres-twitter* (chiffres de juin 2013) consulté le 23 juin 2014. Les 3 V classiques du *Big Data* ne sont cependant pas la seule difficulté que doit affronter le TAL pour accéder au sens des textes : il doit en effet faire face à des caractéristiques intrinsèques des langues naturelles que nous précisons maintenant.

2.2 Caractéristiques des données textuelles

L'accès au sens des données textuelles, au-delà de toute idée de volume, est rendu difficile par la nature textuelle même de ces données. Celles-ci sont d'une part non structurées (ou faiblement structurées dans le cas de pages au format HTML), se présentant uniquement sous la forme d'une succession d'unités lexicales. Contrairement à de nombreuses données impliquées dans le monde du *Big Data*, il n'existe donc pas pour elles de structure *a priori*, porteuse d'une sémantique facilitant leur interprétation, comme dans le cadre d'une base de données relationnelle par exemple.

Par ailleurs, d'autres caractéristiques *quasi* transparentes pour l'humain rendent leur traitement par une machine extrêmement problématique, laissant d'ailleurs le premier parfois perplexe face à l'incompétence de la seconde. Parmi elles, l'ambiguïté, présente à tous les niveaux de l'analyse linguistique, rend les traitements automatiques ardus. Elle peut par exemple être lexicale (*président* pouvant être un nom ou un verbe), syntaxique (structure hiérarchique ambiguë dans *La petite brise la glace* où *brise* peut jouer le rôle de verbe ou de nom sujet) ou encore sémantique (découlant de cas d'homonymie (*avocat* : fruit *versus* personne) ou de polysémie (*porte* : objet *versus* ouverture), voire de l'interprétation de la portée de quantificateurs).

L'implicite est une autre propriété intrinsèque du langage qui justifie d'ailleurs la fréquence des ambiguïtés signalées précédemment, résolubles grâce à lui. Le langage est en effet une interaction entre des personnes qui possèdent des connaissances sur le monde – que celles-ci soient encyclopédiques, de sens commun ou autres – qui leur permettent d'interpréter les énoncés. Ce sont également ces connaissances qui rendent les emplois métaphoriques ou métonymiques compréhensibles.

La possibilité de formuler de façons différentes une même idée est aussi une caractéristique importante du langage naturel. Même si elle n'est pas forcément source de difficulté d'interprétation pour un locuteur natif, cette propriété complexifie la tâche des traitements automatiques dans lesquels il est nécessaire de compter le nombre d'occurrences d'un même concept. Divers types de variation peuvent ainsi recouvrir des concepts très proches : au niveau d'un mot simple ou composé, la variation peut être graphique (changement de casse…), morphologique (mise au pluriel…), syntaxique (modification

de la structure interne du mot), sémantique (remplacement d'un mot par un synonyme…). La reformulation peut aussi s'étudier au niveau d'une phrase paraphrasée par une autre.

Les volumes et les formes plus ou moins dégradées des textes, alliés aux caractéristiques du langage évoquées ci-dessus, rendent les analyses linguistiques fines rapides des énoncés textuels concrètement impossibles, et l'accès au sens qu'ils expriment réellement très difficile. Lorsqu'on se place en domaine ouvert, c'est-à-dire lorsque l'on ne se restreint pas à un domaine spécialisé mais que l'on traite des données textuelles abordant des sujets variés, les ambiguïtés se multiplient et des sources de connaissance telles que des ontologies structurant des concepts ne sont plus utilisables. Un changement de paradigme du TAL s'est donc opéré dès les années 90 face à la disponibilité de volumes de textes déjà très conséquents, avec la perte de vitesse des travaux dédiés à la compréhension fine des énoncés au profit d'études impliquant des représentations beaucoup moins élaborées du sens, suffisantes cependant pour un nombre important d'applications. Nous allons nous pencher sur ces représentations communément utilisées en TAL et sur leurs modes d'élaboration et d'exploitation habituels.

3. Représentations des données textuelles

Les modélisations fines du sens des énoncés, qu'elles reposent sur une vision syntaxico-logique ou sur une optique plus proche de l'intelligence artificielle (IA), ont donc peu à peu laissé la part belle à des représentations des données textuelles à grain beaucoup plus grossier, variant, selon les besoins, du très gros grain sans véritable notion de sémantique à des représentations que l'on pourrait qualifier « du sens utile » pour l'objectif choisi. Les travaux de cette lignée fortement prépondérante en TAL depuis les années 90 sont des études où le sens d'un texte est exprimé par une collection de mots endogènes qui, ensemble, modélisent son contenu ; où le sens d'un mot est exprimé par les mots qui apparaissent près de lui ; où l'expertise est remplacée par les données et où l'on ne cherche plus en fait vraiment à représenter le sens en tant que tel mais à extraire des mots, phrases ou textes des éléments de représentation utiles, exploités ensuite par des algorithmes d'apprentissage artificiel.

Plus précisément, certains travaux de TAL se focalisent sur des représentations sous forme de n-grammes de caractères (séquences de *n* caractères) contenus dans des données textuelles, d'autres sur des n-grammes de mots. Que ce soit pour des mots – en s'intéressant au voisinage de leurs occurrences dans des *corpora* –, des phrases ou des textes – en particulier dans des applications de recherche d'information (RI) [16] ou de catégorisation de textes – les mots mêmes des données textuelles peuvent être retenus comme représentation, très fréquemment au sein d'un sac de mots où toute notion de séquentialité du texte se dissout. Une sélection peut être effectuée sur les mots conservés dans la représentation, fondée sur leur « type » (nom (N), verbe (V), mot composé, entité nommée (nom de lieu, de personne…)) ou encore sur leur saillance, celle-ci s'appuyant très fortement sur un comptage d'occurrences. Les éléments de représentation gardés sont potentiellement associés à une pondération visant à refléter leur importance au sein du texte ou de la phrase considéré(e), par exemple de type *tf*idf* (fréquence du terme x fréquence documentaire inverse, où la fréquence documentaire correspond au nombre de documents/phrases dans lesquel(le)s le mot considéré apparaît) pour pointer les mots fréquents dans l'unité textuelle considérée mais rares par ailleurs. Des représentations sous forme de vecteurs de mots, éventuellement pondérés, peuvent alors être comparées à l'aide de différentes mesures de similarité (ensemblistes, géométriques comme la mesure du cosinus des angles formés par les vecteurs) censées refléter une similarité sémantique.

Pour construire ces représentations, les travaux de TAL appliquent un certain nombre d'outils aux données. Parmi ceux-ci, on peut citer les segmenteurs qui découpent le texte en phrases et en mots ; les

étiqueteurs morphosyntaxiques capables d'affecter aux mots leur étiquette catégorielle ; les analyseurs morphologiques, les racineurs et les lemmatiseurs qui permettent de ramener les mots d'une « famille » à une même forme de base ; les extracteurs de termes, les extracteurs et catégoriseurs d'entités nommées, les analyseurs en dépendance… Les productions de ces outils peuvent servir à sélectionner les éléments de représentation d'une unité textuelle (par exemple les N et V les plus fréquents d'un texte), mais aussi participer elles-mêmes aux représentations. Beaucoup de tels outils sont disponibles gratuitement pour de nombreuses langues, et sont de plus en plus performants sur des textes dégradés issus de systèmes de reconnaissance.

Comme nous l'avons mentionné précédemment, le TAL empirique, en particulier à l'ère du *Big Data*, fait un usage massif de techniques d'apprentissage artificiel [17, 19] pour faire émerger des connaissances des *corpora* de données textuelles. L'apprentissage artificiel est une branche de l'IA qui étudie comment on peut écrire des programmes qui s'améliorent en se confrontant aux données. Les méthodes traditionnellement utilisées en TAL relèvent tant de l'apprentissage supervisé, dans lequel on dispose de données étiquetées (c'est-à-dire pour lesquelles le résultat visé par l'apprentissage est fourni par un expert) que de l'apprentissage non supervisé dans lequel ce n'est pas le cas et où les méthodes tentent par exemple de regrouper des éléments qui se ressemblent (tâche de *clustering* dans laquelle on peut ou non connaître à l'avance le nombre de regroupements (*clusters*) à produire), voire de l'apprentissage semi-supervisé dans lequel le nombre de données étiquetées est réduit. Les représentations de données textuelles peuvent être apprises ; les outils décrits au paragraphe précédent peuvent aussi l'être ; des *clusters* de textes, phrases ou mots proches au niveau de leur sens peuvent être calculés automatiquement à partir de leurs représentations ; enfin des connaissances peuvent être acquises en étudiant des masses de données textuelles et en en faisant émerger des régularités. On peut d'ailleurs considérer que le chercheur en TAL à l'heure du *Big Data* se transforme de plus en plus en un scientifique des données ayant à sa disposition un panel de modes de représentation, de techniques d'apprentissage, de mesures de similarité, de méthodes de visualisation des données, et devant être capable de choisir la méthodologie permettant de faire au mieux sourdre des données l'information souhaitée.

La prise de conscience de l'importance de savoir accéder à l'information contenue dans les données textuelles a cru fortement au cours de ces dernières années. La volonté d'exploiter les connaissances présentes dans des données diverses a conduit au développement de nombreux travaux spécifiques aux données moins habituelles, souvent en adaptant des méthodes issues du TAL « standard », mais aussi de recherches manipulant conjointement des données de types différents. Il est donc difficile de présenter l'ensemble des travaux actuels du TAL à cette époque du *Big Data*. Nous avons par conséquent choisi de nous focaliser sur certains d'entre eux que nous jugeons particulièrement spécifiques de l'aspect « données massives ». Nous les organisons en deux familles que nous décrivons dans les deux sections suivantes. La première concerne l'aide que le TAL peut apporter à un utilisateur face au gigantisme des volumes de données à sa disposition ; la seconde rassemble des travaux qui tentent de tirer au mieux profit de la redondance présente dans les données massives.

4. Faire face au déluge de données

Confronté à une avalanche de données textuelles, un utilisateur peut avoir besoin d'une aide pour se retrouver dans l'ensemble d'informations à sa disposition. De nombreuses recherches de TAL tentent de répondre à cette nécessité depuis plusieurs années. Nous en abordons ici certaines, ayant parfaitement conscience d'avoir effectué une sélection sévère et également de présenter les travaux retenus sous l'angle qui sert notre propos global d'assistance à l'usager ; cette sélection donne cependant une vue assez claire de certaines problématiques phares du TAL à l'ère du *Big Data*. Nous

nous intéressons en section 4.1 au résumé automatique de texte mono ou multi-documents dont l'objectif est de faire ressortir les idées-force d'un (paquet de) document(s), par exemple pour savoir s'il est intéressant à lire *in extenso* ou non. Plus globalement, structurer, en se fondant sur son contenu langagier, la masse d'informations disponible et proposer à l'utilisateur une ou plusieurs grille(s) de lecture ou de navigation peut aussi lui éviter une désorientation ; c'est ce sujet que nous abordons en section 4.2. Avoir rapidement une vue de la perception que les internautes ont d'une personne ou d'un produit est également un besoin fort qui a conduit à une profusion de recherches en fouille d'opinion que nous résumons en section 4.3. Enfin, la volonté d'exploiter les données textuelles de la même façon que des données structurées dans des cadres d'aide à la décision a fait émerger le domaine du *Text analytics* que nous décrivons brièvement pour terminer.

4.1 Résumé automatique mono ou multi-documents

Initié par des premiers travaux dès la fin des années 50, le domaine du résumé automatique de textes a connu un essor très conséquent au milieu des années 90 qui ne se dément pas depuis [5]. Les travaux de ce domaine, qui consistent en règle générale, sous la contrainte d'une taille plus ou moins précise, à produire un texte qui contient l'information importante du ou des textes initiaux, peuvent, même s'ils partagent un nombre important d'aspects méthodologiques, se distinguer selon divers axes. Certaines recherches génèrent un texte contenant les idées du document initial (résumé par abstraction) tandis que d'autres, beaucoup plus nombreuses, procèdent par extraction de phrases du document, éventuellement avec un post-traitement visant à rendre le résultat plus lisible. Le point de départ de la création du résumé peut être le texte initial lui-même ou un besoin exprimé sous forme d'une requête. Les travaux peuvent concerner des méthodes de résumé mono ou multi-documents. Enfin, la méthodologie globale peut avoir un fort ancrage linguistique, reposer sur de l'apprentissage supervisé des caractéristiques des phrases à conserver, ou exploiter fortement des représentations proches de la RI.

Nous nous focalisons ici sur les études qui procèdent par extraction de phrases d'un ou plusieurs documents. Afin de déterminer les phrases saillantes à conserver, un score est calculé pour chacune d'entre elles en combinant divers indicateurs. Parmi les traits exploités figurent la saillance des mots (fondée sur un score *tf* ou *tf***idf*), la position de la phrase dans le texte, sa longueur, la présence ou non de certains mots-clés, de marqueurs de structure, d'entités nommées… La prise en compte de la seule saillance des phrases, dans le cas du résumé mono-document mais encore plus dans le cas du multi-documents, ne suffit toutefois pas : il faut aussi que les phrases retenues ne soient pas redondantes. La mesure MMR (*maximal marginal relevance*) [8] établit ainsi pour chaque phrase un score combinant sa saillance à sa similarité (par exemple évaluée via un cosinus entre vecteurs de mots) avec les phrases déjà dans le résumé.

Actuellement trois familles d'approches sont prépondérantes et se distinguent en particulier par leur façon de gérer la redondance. Dans l'extraction de phrases fondée centroïde [23], le score d'une phrase dépend entre autres de sa centralité par rapport au thème global des documents à résumer – établie par les mots d'un document centroïde de l'ensemble qu'elle contient – et de sa redondance par rapport aux phrases déjà conservées. Dans l'extraction de phrases fondée graphe, chaque phrase est un nœud et on décide de l'importance d'un nœud en prenant en compte l'information globale calculée récursivement à partir du graphe entier. Dans l'algorithme TextRank [18], chaque phrase est liée aux phrases avec lesquelles elle partage du vocabulaire par un arc pondéré. Un score initial est attribué à chaque nœud et un algorithme itère jusqu'à convergence en recalculant le score des nœuds en fonction du score de ceux qui leur sont liés et du poids des arcs. Les phrases à scores les plus élevés sont conservées. Enfin, dans la production de résumé fondée programmation par contraintes, le but est de maximiser une

fonction-objectif globale ; par exemple dans [7], la sélection de phrases que doit faire le solveur de contraintes doit optimiser la somme des valeurs de concepts – bigrammes informatifs de mots pondérés par leur fréquence – que le résumé contient ; un résumé est donc d'autant meilleur qu'il couvre un nombre élevé de concepts différents, d'où sa non-redondance.

Les études sur le résumé de transcriptions de la parole sont beaucoup moins nombreuses. Certaines concernent toutefois le résumé de réunions ou celui de vidéos et nécessitent parfois la détermination des fins de phrases au sein des groupes de souffle.

L'évaluation des résumés reste un problème épineux, l'accord entre humains sur la définition même d'un bon résumé étant difficile à obtenir. L'évaluation automatique est fondée sur des mesures telles que ROUGE qui calcule globalement la proportion de n-grammes partagés par le résumé proposé et un ou plusieurs résumés de référence rédigés par des humains. En dépit du travail réalisé, les résumés par extraction souffrent encore de limitations : d'un point de vue linguistique, la qualité et la cohésion textuelles restent par exemple encore des problèmes ouverts.

4.2 Structuration et navigation

Outre leur volume et leur hétérogénéité, les données textuelles actuellement disponibles manquent aussi de structure explicite et sont, pour la plupart, déconnectées les unes des autres : il n'existe en effet pas entre elles de liens indiquant qu'elles portent sur le même sujet, que l'une est la continuation de ce qui est abordé dans l'autre, voire une opinion exprimée sur ce qu'elle contient… Un effort d'organisation de cette masse de données et de création de possibilités de navigation éclairée dans son contenu est donc nécessaire, tant pour des utilisations individuelles que dans le cadre d'applications professionnelles, pour éviter une désorientation et tirer au mieux profit de l'information présente.

Des efforts en ce sens ont été réalisés en RI *a posteriori* de l'obtention des résultats d'une requête, une visualisation organisée des résultats fondée sur leur contenu [11] donnant à l'utilisateur une perception plus abstraite de leur typologie. On peut également signaler le développement du domaine de la question-réponse [12] qui permet à un individu d'obtenir une réponse à une interrogation précise plutôt que de devoir effectuer la recherche au sein des documents renvoyés. Toujours *a posteriori* d'une requête, des organisations plus informatives des documents retournés sont maintenant proposées, nécessitant pour ce faire des représentations des données mêlant le grain assez grossier des sacs de mots « standards » à des analyses linguistiques plus fines. [26] présente ainsi la création automatique de chronologies événementielles thématiques au sein de dépêches AFP. À partir d'une requête assez générale (telle que « le printemps arabe ») adressée à un moteur de recherche indexant ces dépêches, un travail de représentation des aspects temporels permet de faire émerger des documents renvoyés les dates qui sont considérées comme saillantes (saillance calculée grâce au score de pertinence fourni par le moteur et par apprentissage supervisé par rapport à des chronologies de référence), et donc de proposer un ordonnancement dans le temps des événements correspondants.

Établir des liens entre documents ou fragments de documents pour offrir des possibilités de navigation dans les données ne nécessite pas forcément de requête initiale. L'établissement de liens fondé sur la comparaison de contenus a été initié par la communauté de l'hypertexte qui visait à enrichir les textes avec des hyperliens [1]. Cette création d'hypertextes a surtout concerné des documents avec une structure assez marquée (*emails*, pages Wikipedia…), ou des collections limitées à nouveau pour naviguer dans les documents-réponses à une question. La recommandation est aussi un domaine typique où des liens entre contenus sont créés. Les propositions des systèmes exploitent souvent le filtrage collaboratif (*i.e.*, les appréciations d'utilisateurs) pour établir des proximités, conjointement à des mesures standards de similarité fondées contenu [4]. Des liens sémantiques informatifs peuvent aussi être recherchés au sein des textes. [26] propose ainsi d'organiser des dépêches AFP au sein d'un

graphe temporel où les articles sont liés par des relations même événement, continuation (conséquence…) et réaction (expression d'une opinion sur le sujet). Ce travail repose sur un apprentissage supervisé (classifieurs relation *versus* pas de relation, relation même événement *versus* continuation…). [9] s'intéresse également à l'établissement de liens fondés contenu mais exploite pour ce faire la transcription de la parole exprimée dans un mois de journaux télévisés de France 2. La prise en compte au sein d'un système de segmentation thématique de textes des mesures de confiance associées aux mots transcrits et de relations sémantiques apprises en corpus permet de pallier les particularités des transcriptions et d'offrir automatiquement un découpage en sujets successifs. Une extraction de quelques mots-clés de chaque sujet, à nouveau en modifiant les mesures traditionnelles de saillance par les mesures de confiance, permet d'une part la création de liens entre les sujets similaires dans différents journaux, mais aussi vers des pages Web apportant des informations complémentaires aux points traités (*cf.* une version du système prenant en compte des fonctionnalités supplémentaires à texmix.irisa.fr).

Comme illustré par ces deux derniers travaux, il devient donc possible, en se fondant sur le seul contenu, quitte à mêler des représentations à grains distincts, d'établir des liens entre des données textuelles hétérogènes, mais également entre des données issues de médias différents *via* la prise en compte de la transcription de la parole par exemple. Ceci soulève toutefois la question de la granularité des documents à laquelle créer les liens (entre documents entiers, entre sources et cibles précises…) mais également des usages de ces nouvelles structurations de l'information.

4.3 Fouille d'opinion

L'essor du Web social (*fora*, blogs, réseaux sociaux…) a conduit à la prolifération de données textuelles au sein desquelles sont exprimées des opinions, des évaluations, des sentiments, données qui sont le sujet d'étude de la fouille d'opinion. Les enjeux de ce domaine sont multiples, entre autres économiques et politiques (conseils pour un achat, stratégies marketing, perception de la mise en place d'une réforme…).

Les données du Web social sont très diverses, en qualité d'écriture mais aussi au regard de l'information d'opinion qu'elles véhiculent : elles peuvent être focalisées sur l'expression d'un avis global sur une seule entité, émettre des opinions sur différents aspects d'une même entité, mentionner des commentaires sur des entités diverses, contenir à la fois des zones où des opinions sont émises et des zones plus factuelles ; enfin certaines s'interprètent seules alors que d'autres se répondent et ne peuvent être comprises que dans le contexte du « fil » auxquelles elles appartiennent.

Trois tâches reposant sur l'analyse des contenus textuels ont été ou sont principalement étudiées en fouille d'opinion [15]. La première consiste à séparer les textes ou portions de textes portant une opinion de ceux ou celles qui n'en contiennent pas, alors que la deuxième vise à attribuer une polarité positive, négative ou neutre à l'opinion exprimée. Ces deux tâches font massivement appel à des méthodes d'apprentissage, apprentissage fréquemment supervisé exploitant des représentations très variées (mots pondérés ou pas, n-grammes de mots ou caractères, étiquettes morphosyntaxiques, ponctuation, position, présence de mots exprimant des sentiments…). Les travaux fondés sur de l'apprentissage non supervisé font usage de lexiques de mots ou de syntagmes polarisés ou de quelques mots amorces, voire de règles et d'heuristiques linguistiques pour aboutir à une classification. Les lexiques peuvent eux-mêmes faire l'objet d'un apprentissage, avec des mots amorces et une expansion *via* une ressource généraliste (*cf. SentiWordNet*), ou un corpus et quelques heuristiques. La troisième tâche concerne la production de résumés d'opinions. La synthèse produite peut être textuelle, des méthodes proches de celles présentées en 4.1 pouvant être adaptées par la prise

en compte de traits relatifs à l'expression d'opinions. La connaissance des éléments constitutifs de l'entité visée peut influer sur la forme du résumé, avec une première phrase générale la concernant dans sa globalité, suivie de phrases pour chacun des aspects référés ; plus qu'un avis moyen, les avis extrêmes peuvent être recherchés ; enfin une connaissance du nombre d'opinions positives et négatives est fréquemment souhaitée. Une formulation graphique du résumé est donc souvent favorisée dans ce domaine, par exemple sous forme de diagrammes circulaires ou à bâtons (*cf*. l'outil *SenticNet*). Dernier point, si la date d'émission des opinions est connue, une visualisation de l'évolution de l'avis sur l'entité peut être produite.

Malgré des avancées très nombreuses, beaucoup de travail reste encore à faire dans le domaine, les éléments présentés restant des sujets d'études. Le passage détermination de la polarité de l'opinion globale émise dans un document des années 2000 à la production des quintuplets (source de l'opinion, entité-cible, aspect-cible de l'entité, opinion, date de celle-ci) décrits par [15], s'il est possible dans certains cas, reste encore fortement non systématique, un ensemble de difficultés de TAL (négation, anaphores, implicite…) bridant entre autres les avancées. De nouveaux champs de recherche émergent aussi, parmi lesquels la contextualisation des opinions émises au sein d'un fil (de *tweets* par exemple) ou encore la détection de faux avis sur les sites de commentaires de produits.

4.4 Text analytics

Avant de conclure cette section, même si cela n'implique pas l'explicitation d'un nouveau domaine de recherche en tant que tel, il convient de souligner la prise de conscience de l'importance des données textuelles au sein des entreprises et l'essor du champ du *Text analytics* qui en découle. Les données textuelles disponibles dans les entreprises concernent tant les documents, rapports, brevets, *emails*, remontées des centres d'appels que les commentaires postés sur le Web social ou les pages Web de leur domaine d'intérêt. Les entreprises souhaitent catégoriser leurs documents, y rechercher de l'information, mais aussi savoir ce que les consommateurs pensent de leurs produits et de ceux de leurs concurrents, détecter des sujets émergents et utiliser conjointement données structurées et non structurées pour des modélisations descriptives et prédictives. Les grands éditeurs de logiciels ainsi que des sociétés de taille plus modeste se sont résolument engagés dans ce domaine ces dernières années.

5. Exploiter la profusion et la redondance

La présence de volumes très conséquents de données textuelles, si elle peut être vue comme une contrainte nécessitant la mise en place d'éléments d'aide pour les utilisateurs, peut aussi être considérée comme une force : la profusion et la redondance d'information peuvent permettre de constituer une connaissance plus complète et plus certaine de nombreux sujets. Nous présentons successivement dans cette section une sélection de trois domaines de recherche qui se fondent sur cette idée : la traduction automatique, le journalisme de données et la vérification par les faits. Le premier, domaine ancien du TAL, a connu un bouleversement dans ses méthodes avec la disponibilité de volumes importants de textes en plusieurs langues ; les deux autres sont des sujets beaucoup plus récents.

5.1 Traduction automatique

Les recherches sur la traduction automatique des langues ont débuté quelques années seulement après l'apparition des premiers calculateurs électroniques [13]. Jusqu'à la fin des années 80, trois approches fondées sur des règles linguistiques de tout type (lexicales, d'analyse morphologique, syntaxique, de génération syntaxique…) ont coexisté : l'approche directe s'intéressant à la traduction d'une langue

source en une langue cible à l'aide d'un dictionnaire et de quelques règles syntaxiques ; le modèle interlangue reposant sur une représentation intermédiaire abstraite impliquant la traduction de la source vers l'interlangue et de l'interlangue vers la cible ; et l'approche par transfert fondée sur une analyse de l'entrée-source pour en produire une représentation, un transfert vers une représentation-cible et une génération en langue cible. Ces systèmes à base de règles étaient principalement utilisés par des institutions et des entreprises (notons aussi l'existence d'outils d'aide à la traduction humaine).

L'avènement d'Internet, la mondialisation des échanges, les nouveaux besoins qui en ont découlé – nécessité pour chacun de comprendre et produire rapidement des textes, des *emails*, de styles et de sujets variés, dans une langue autre que la sienne – ont, dès les années 90, totalement bouleversé le paysage de la traduction automatique. L'approche à base de règles a vu son hégémonie balayée par l'émergence de celle fondée corpus, rendue possible par la disponibilité de grands volumes de textes en diverses langues.

Au sein de cette approche, la traduction automatique statistique (TAS) [2], initiée au début des années 90 par une équipe d'IBM [3], est devenue le courant majoritaire. Dans ce cadre, traduire une phrase *f* en langue source en une phrase *e* en langue cible revient à chercher la phrase notée *e** qui maximise la probabilité de *e* sachant *f* (*P(e|f)*). En appliquant la règle de Bayes, on exprime plus clairement le désir de se concentrer sur les phrases de la langue cible correctes et traductions de *f* :

$$e^* = argmax_e\ P(e|f) = argmax_e\ P(f|e)\ P(e).$$

Chercher à maximiser le produit revient à décomposer le problème en deux :

- d'une part développer un *modèle de traduction* garantissant que *P(f|e)* est élevée pour toute phrase en langue cible, grammaticale ou pas, appariée à *f* ; ce modèle est estimé sur de grands *corpora* bilingues parallèles alignés au niveau de la phrase, c'est-à-dire des ensembles de textes en deux langues qui sont des traductions (le *Hansard*, débats parlementaires canadiens en Français et en Anglais (environ 20 millions de mots par langue), ou *Europarl*, débats parlementaires européens en 21 langues (environ 60 millions de mots par langue) sont souvent utilisés) ;
- d'autre part développer un *modèle de* la *langue* cible, associant des valeurs de probabilités élevées aux phrases grammaticales et fournissant la valeur *P(e)* indépendamment de la phrase à traduire ; de tels modèles sont couramment établis à partir de *corpora* monolingues (estimation des probabilités de n-grammes).

Dans les travaux de Brown *et al.* [3], le modèle de traduction est vu comme un modèle d'alignements de mots. Cinq modèles, dits modèles IBM, ont été proposés dont certains sont à la base d'un grand nombre de modèles à base de segments (groupes de mots) qui sont, depuis le début des années 2000, devenus le standard en TAS. Dans ce nouveau cadre, un système de traduction fonde globalement sa décision en s'appuyant sur des valeurs numériques issues du modèle de traduction, qui évalue la qualité de l'appariement *f-e* établi grâce aux choix d'une segmentation de *f* et à celui, pour chacun des segments, d'un équivalent en langue cible, d'un modèle de distorsion qui évalue la plausibilité du ré-ordonnancement des segments induit par l'appariement retenu (gestion des ordres différents entre langues source et cible), et du modèle de langue qui évalue la qualité de la phrase cible formée. La performance d'un système est évaluée selon diverses métriques, parmi lesquelles BLEU qui repose sur une comparaison des n-grammes présents dans la traduction et dans une ou des traductions références.

Une seconde approche fondée corpus s'est aussi développée, sans être aussi massive que la TAS : la traduction à base d'exemples [20] qui utilise un ensemble de phrases déjà traduites, éventuellement abstraites pour accroître la généralisation. Cette approche repose sur deux composants : un algorithme

de mise en correspondance de la phrase source avec les exemples, c'est-à-dire l'identification des parties de la phrase appariables avec des exemples, et un algorithme de recombinaison des éléments traduits pour obtenir une phrase en langue cible.

La traduction automatique a été et est aussi appliquée à la parole, en particulier ces dernières années sans contrainte de domaine dans un cadre statistique unifié. Certains travaux explorent la traduction *a posteriori* de la transcription, reponctuant les transcriptions, supprimant les disfluences… pour la faciliter, tandis que d'autres se focalisent par exemple sur la traduction des sorties intermédiaires du système de reconnaissance et mêlent les scores des deux systèmes pour ordonner les traductions.

5.2 Journalisme de données

Le journalisme de données (JD) est un terme récent qui décrit une pratique journalistique fondée sur la collecte, le filtrage, la combinaison et l'analyse de grands volumes de données afin d'en faire émerger une histoire digne d'être racontée. Ses principes-clés impliquent la découverte dans les données de faits intéressants, la mise en évidence de tendances cachées, la compilation d'ensembles de données accessibles par ordinateur mis à disposition du public, et leur visualisation appropriée. Le but est d'expliquer des histoires potentiellement complexes en s'appuyant sur des graphiques clairs, qui peuvent être interactifs et personnalisables. Le domaine est conceptuellement et sur un plan pratique assez proche de l'intelligence économique (*business intelligence* ou *analytics*, *cf.* section 4.4) mais la méthodologie employée n'est plus à destination des entreprises mais des médias et du grand public. Il est assez communément admis que l'une des premières formulations de ce qui est appelé JD a été produite par A. Holotavy en 2006 [14]. La pratique a cependant aussi des liens forts avec des démarches plus anciennes telles que le journalisme d'investigation. Un guide du JD, initié en 2011 et regroupant ses fondamentaux, est disponible à www.datajournalismhandbook.org.

Les données utilisées dans le cadre du JD proviennent de collectivités, de services publics (*cf.* www.data.gouv.fr), du Web, voire sont parfois confidentielles (*cf.* WikiLeaks). Leur abondance donne accès à des informations complètes, le croisement de plusieurs sources permettant, lui, d'assurer leur fiabilité. Si les données exploitées en JD de manière automatisée sont, pour une grande part, actuellement structurées, la connaissance présente souvent uniquement dans la masse de données textuelles existante doit aussi être prise en compte. Le TAL et la RI offrent des moyens d'accès à cette connaissance qui ont déjà fait leurs preuves et dont on peut prévoir qu'ils vont être amenés à être de plus en plus usités. Les trois exemples ci-dessous utilisent de façon plus ou moins poussée leurs potentialités.

J. Stray et J. Burges [25] ont, en 2010, travaillé sur les résumés textuels des rapports d'enquêtes de l'*US Army* concernant la guerre en Iraq entre 2004 et 2009 dévoilés par WikiLeaks. Ils proposent une visualisation thématique des documents de décembre 2006 obtenue grâce à une représentation des résumés sous forme de vecteurs de mots avec pondération *tf*idf* et à une comparaison par mesure cosinus. Les mots les plus caractéristiques de chaque document permettent de faire apparaître sur les *clusters* du graphe formé leurs mots les plus saillants.

J. Véronis propose, lui, une analyse linguistique plus fine de discours politiques, en particulier dans le cadre de la campagne présidentielle 2012 (*cf.* son *Observatoire des discours* sur le site du Monde), se focalisant sur l'utilisation des pronoms (en particulier du « je »), mais aussi sur les thèmes abordés caractérisés par la présence de mots-clés, les entités nommées employées… Des commentaires sont proposés sur le blog *Politicosphère* du journal, par exemple sur l'emploi de l'expression « Moi président » par F. Hollande [27].

Le travail de recherche de X. Tannier [26] propose un outil d'identification automatique des relations d'alliance et d'opposition entre pays sur un sujet donné, montrant les potentialités du TAL pour automatiser effectivement l'accès au contenu textuel dans un cadre de JD. Les visualisations offertes (courbe, graphe, carte) montrent l'évolution des relations dans le temps. Ce travail repose entre autres sur une extraction automatique des associations entre pays et capitales, villes ou personnes, un lexique de 110 déclencheurs de relations (*argue*, *criticism*…), une normalisation des dates, l'apprentissage d'un classifieur séparant les phrases porteuses ou non d'une relation positive ou négative entre pays, et un moteur de recherche indexant celles contenant une relation.

5.3 Fact-checking

Le *fact-checking* (FC, vérification par les faits) est la démarche qui a pour objet de vérifier la véracité d'affirmations faites lors de discours publics ou au sein de documents en les comparant à des faits présents dans diverses sources d'informations sûres. L'objectif n'est pas seulement de démontrer que l'affirmation est correcte ou erronée mais de souligner omission, lecture à sens unique ou autre déformation de l'information. Popularisé par des sites tels que *PolitiFac* ou par des émissions télévisuelles et radiophoniques telles que « Le vrai du faux » sur France-Info, le FC, qui a pour cible favorite les déclarations des politiciens, est à l'heure actuelle essentiellement manuel : c'est en effet ainsi, en recoupant plusieurs bases de faits dont ils connaissent l'existence et la fiabilité, que les *fact-checkers* contrôlent les affirmations.

À l'instar d'autres pans de l'informatique, en particulier pour ce qui est des données (semi-)structurées, le TAL a cependant de nombreux outils et méthodes à offrir au FC pour tirer profit des informations présentes dans les données textuelles et pour vérifier celles qui s'y trouvent. Ainsi, les travaux en extraction d'informations peuvent extraire un fait et ses attributs ; des moyens pour faire émerger par apprentissage supervisé ou non des relations entre éléments sont aussi disponibles ; le domaine de la question-réponse peut être sollicité ; la RI offre des méthodes de comparaison entre contenus aptes à rechercher et recouper des données ; il est possible, comme on l'a vu, de retracer le fil d'un événement… Parmi les travaux actuels qui sont concrètement les plus proches du domaine du FC, certains s'intéressent aux éléments linguistiques permettant d'établir la croyance du rédacteur en la véracité du fait qu'il mentionne, employant des méthodes soit à base d'apprentissage supervisé, soit fondées règles et expertise humaines [24]. Un autre ensemble d'études concerne la recherche des faits vrais parmi les multiples sources qui attestent pour un même fait des valeurs différentes. L'objectif est de tirer profit de la redondance de l'information pour démêler les candidats-faits vrais des faux. Plusieurs algorithmes reposent sur un vote itératif transitif [28] : chaque candidat-fait reçoit un score de crédibilité initial ; la fiabilité des sources est alors calculée à partir des scores des candidats qu'elles mentionnent ; la crédibilité des candidats est ensuite réévaluée grâce à la fiabilité des sources ; l'algorithme itère jusqu'à convergence. [21] utilise l'objectivité des sources pour déterminer leur fiabilité, objectivité détectée par apprentissage sur la base de la présence ou non de traits linguistiques tels que des mots issus de lexiques d'opinion. [22] traite le sujet sous la forme d'un problème d'inférence probabiliste. Les perspectives quant à l'emploi du TAL dans ce contexte sont donc encore très importantes.

6. Conclusion

Comme illustré dans ce chapitre, le *Big Data* n'a pas été source d'une révolution du TAL, qui avait subi une mutation profonde dès la fin des années 80 et, entre autres, l'avènement d'Internet ; il a toutefois conduit le domaine à se confronter à des données textuelles de qualités très diverses, à revisiter des applications existantes et à en proposer de nouvelles, soit pour faire face au déluge de

données, soit parce que ce déluge est lui-même porteur de solutions. Sur un plan plus technique, il a peu à peu transformé le chercheur en TAL, proche de la langue, en un scientifique des données capable de choisir la bonne représentation – souvent peu élaborée linguistiquement et cognitivement – et la bonne méthode d'apprentissage pour faire émerger l'information souhaitée, voire la bonne méthode de visualisation pour la présenter. Si les grandes conférences du TAL reflètent fortement cette tendance – on note même l'émergence du paradigme de la « découverte » dans lequel on laisse les données exprimer elles-mêmes la connaissance qu'elles contiennent –, les représentations mêlant un grain grossier à un grain linguistiquement plus fin semblent toutefois peu à peu revenir sur le devant de la scène. Au-delà de ces constatations, la retombée la plus forte du *Big Data* sur le TAL est peut-être la prise de conscience par tous de l'importance de disposer de moyens d'accès au contenu sémantique des données textuelles, et une attente d'applications efficaces et capables d'expliquer les actions et décisions de leurs méthodologies sous-jacentes.

Bibliographie

[1] AGOSTI (M.) et ALLAN (J.), *Special issue on methods and tools for the automatic construction of hypertext*, *Information Processing and Retrieval*, Vol. 33(2), 1997.

[2] ALLAUZEN (A.) et YVON (F.), *Méthodes statistiques pour la traduction automatique*, dans GAUSSIER (É.) et YVON (F.), *Modèles statistiques pour l'accès à l'information textuelle*, Paris, Hermès, chapitre 7, 2011, p. 271-356.

[3] BROWN (P.F.), COCKE (J.), DELLA PIETRA (S.A.), DELLA PIETRA (V.J.), JELINEK (F.), LAFFERTY (J.D.), MERCER (R.L.) et ROOSSIN (P.S.), *A statistical approach to machine translation*, dans *Computational Linguistics*, Vol. 16(2), 1990, p. 79-85.

[4] CLAYPOOL (M.), GOKHALE (A.), MIRANDA (T.), MURNIKOV (P.), NETES (D.) et SARTIN (M.), *Combining content-based and collaborative filters in an online newspaper*, dans Actes de ACM SIGIR Workshop on recommender systems: algorithms and evaluation, 1999.

[5] DAS (D.) et MARTINS (A.F.T.), *A survey on automatic text summarization*, Rapport technique, Literature Survey for the Language and Statistics II course at Carnegie Mellon University, 2007.

[6] FAIRON (C.), KLEIN (J.) et PAUMIER (S.), *Le langage SMS. Étude d'un corpus informatisé à partir de l'enquête « Faites don de vos SMS à la science »*, Louvain-la-Neuve, Presses universitaires de Louvain, 2006.

[7] GILLICK (D.) et FAVRE (B.), *A scalable model for summarization*, dans Actes de NAACL HLT Workshop on integer linear programming for natural language processing, 2009, p. 10-18.

[8] GOLDSTEIN (J.) et CARBONELL (J.), *Summarization : (1) using MMR for diversity-based reranking and (2) evaluating summaries*, dans Actes de Workshop on TIPSTER text program: Phase III, 1998, p. 181-195.

[9] GRAVIER (G), GUINAUDEAU (C.), LECORVÉ (G.) et SÉBILLOT (P.), *Exploiting speech for automatic TV delinearization: From streams to cross-media semantic navigation*, dans *Eurasip Journal on Image and Video Processing*, Vol. 2011, 2011.

[10] HALPER (F.), KAUFMAN (M.) et KIRSH (D.), *Text analytics: The Hurwitz victory index report*, Hurwitz & Associates, 2013.

[11] HEARST (M.A.) et PEDERSEN (P.), *Reexamining the cluster hypothesis: Scatter/gather on retrieval results*, dans Actes de 19th Annual international ACM/SIGIR conference, 1996, p. 76-84.

[12] HIRSCHMAN (L.) et GAIZAUSKAS (R.), *Natural language question answering*, dans *Natural Language Engineering*, Vol. 7(4), 2001, p. 275-300.

[13] HUTCHINS (J.W.), *Machine translation: history of research and applications*, dans Chan Sin-wai, *Routledge Encyclopedia of Translation Technology*, chapitre 6, 2015, à paraître.

[14] HOLOTAVI (A.) (2006, 6 sep.), *A fundamental way newspaper sites need to change*, sur http://www.holovaty.com/writing/fundamental-change/. Consulté le 9 juil. 2014.

[15] LIU (B.), *Sentiment analysis and opinion mining*, Morgan & Claypool publishers, 2012.

[16] MANNING (C.D.), RAGHAVAN (P.) et SCHÜTZE (H.), *Introduction to information retrieval*, Cambridge University Press, 2008.

[17] MANNING (C.D.) et SCHÜTZE (H.), *Foundations of statistical natural language processing*, Cambridge, Massachusetts, The MIT Press, 1999.

[18] MIHALCEA (R.), *Graph-based ranking algorithms for sentence extraction, applied to text summarization*, dans Actes de 42nd Annual meeting of the association for computational linguistics on interactive poster and demonstration sessions, 2004.

[19] MITCHELL (T.), *Machine learning*, McGraw Hill, 1997.

[20] NAGAO (M.), *A framework of a mechanical translation between Japanese and English by analogy principle*, dans Elithorn (A.) et Banerji (R.), *Artificial and Human Intelligence*, Amsterdam, North-Holland Publishing Company, chapitre 11, 1984, p. 173-180.

[21] NAKASHOLE (N.) et MITCHELL (T.M.), *Language-aware truth assessment of fact candidates*, dans Actes de 52nd Annual meeting of the association for computational linguistics, 2014, p. 1009-1019.

[22] PASTERNACK (J.) et ROTH (D.), *Latent credibility analysis*, dans Actes de 22nd International World Wide Web conference, 2013, p. 1009-1020.

[23] RADEV (D.R.), BLAIR-GOLDENSOHN (S.) et ZHANG (Z.), *Experiments in single and multi-document summarization using MEAD*, dans Actes de 1st Document understanding conference, 2001.

[24] SAURÍ (R.) et PUSTEJOVSKY (J.), *Are you sure that this happened? Assessing the factuality degree of events in text*, dans *Computational Linguistics*, Vol. 38(2), 2012, p. 261-299.

[25] STRAY (J.) et BURGES (J.) (2010, 10 déc.), *A full-text visualization of the Iraq War Logs*, sur http://jonathanstray.com/a-full-text-visualization-of-the-iraq-war-logs. Consulté le 9 juil. 2014.

[26] TANNIER (X.), *Traitement des événements et ciblage d'information*, Habilitation à diriger des recherches, Université Paris Sud, 2014.

[27] VÉRONIS (J.) (2012, 4 mai), *Moi, François Hollande*, sur http://politicosphere.blog.lemonde.fr/2012/05/04/moi-francois-hollande/. Consulté le 9 juil. 2014.

[28] YIN (X.), HAN (J.) et Yu (P.S.), *Truth discovery with multiple conflicting information providers on the Web*, dans Actes de 13th International conference on knowledge discovery and data mining, short paper, 2007, p. 1048-1052.