Le traitement automatique des langues pour les sciences sociales

Quelques éléments de réflexion à partir d'expériences récentes

Thierry Poibeau

Dans Réseaux 2014/6 n° 188 , pages 25 à 51
Éditions La Découverte

ISSN 0751-7971
ISBN 9782707183347
DOI 10.3917/res.188.0025
Date de mise en ligne : 25/02/2015

LE TRAITEMENT AUTOMATIQUE DES LANGUES POUR LES SCIENCES SOCIALES

Quelques éléments de réflexion à partir d'expériences récentes

Thierry POIBEAU

DOI: 10.3917/res.188.0025

Ce numéro de la revue *Réseaux* rappelle clairement les enjeux actuels en sciences sociales : ce domaine de recherche est passé brusquement d'une situation de pénurie de données ou, du moins, d'une situation où les données étaient difficiles à assembler, à une situation où les données sont apparemment massives, disponibles et faciles d'accès. Mais, loin de l'Eldorado promis par les hérauts du *big data*, cette masse de données n'est pas sans poser problème : la plupart du temps, les données ne sont pas directement exploitables, elles doivent être triées, filtrées, organisées ; elles reflètent des points de vue particuliers qui ne sont pas obligatoirement ceux visés par le chercheur en sciences sociales ; enfin, elles peuvent être partielles ou biaisées. Nous garderons ces difficultés en tête lors de cette étude, mais nous nous focaliserons surtout sur le cas des données textuelles : celles-ci constituent une source incomparable de connaissances, mais offrent dans le même temps les plus grandes difficultés d'accès.

En effet, comme chacun le sait, un texte ne saurait être assimilé à une masse de connaissances directement exploitable par la machine. Il faut dans un premier temps prévoir des traitements complexes pour identifier l'information pertinente, la normaliser, la catégoriser et éventuellement la mettre en contexte. Alors seulement l'ordinateur ou l'expert sera capable d'en tirer parti pour mener à bien ses analyses. Mais comment procéder pour extraire l'information pertinente de la masse textuelle ? Quels outils utiliser ? Pour quelle pertinence ? Ces questions sont ouvertes et n'ont pas de réponse immédiate et évidente : cet article présentera un aperçu des techniques et des possibilités actuelles.

Plusieurs études ont pointé la frustration des chercheurs en sciences sociales face à ce problème : les textes sont effectivement là, présents et disponibles sur la Toile, mais leur exploitation reste difficile[1]. Elle exige la collaboration

1. Voici par exemple un témoignage de l'équipe du médialab de Sciences Po : « Qualitative researchers […] arrive at the médialab bringing rich data and longing to explore them. Their problem is that qualitative data cannot be easily fed into network analysis tools. Quantitative data can have many different forms (from a video recording to the very memory of the researcher), but they are often stored in a textual format (i.e. interviews transcriptions, field notes or archive documents…). The question therefore becomes: how can texts be explored quali-quantitatively? Or, more pragmatically, how can texts be turned into networks? » (Venturini et Guido, 2012).

de spécialistes de différents horizons, capables de traiter les données, de fournir les outils pour extraire l'information pertinente et d'ajuster de manière collaborative les traitements. Idéalement, l'exploitation des données disponibles sur la Toile dans le domaine des sciences humaines et sociales nécessiterait la mise en place d'équipes pluridisciplinaires, mais ceci reste rare et difficile, du fait de la spécialisation de plus en plus grande des recherches et des intérêts divergents des différentes disciplines. Il est en effet difficile de définir un projet de recherche à la fois innovant dans le domaine du traitement de l'information et dans le domaine des sciences sociales. Il est alors nécessaire d'avoir recours à des unités de service, qui pourront par exemple fournir des traitements standard aux chercheurs en sciences sociales en utilisant une combinaison d'outils existants[2].

Nous laisserons là ces considérations pratiques pour donner plus prosaïquement un exposé rapide des techniques et des enjeux du domaine. Nous présentons dans la section 2 un aperçu de deux grands types d'analyse visant d'une part l'extraction d'information factuelle à partir de textes et d'autre part l'émergence de modules d'analyse d'informations subjectives, comme les opinions et les sentiments. Nous examinons ensuite dans la section 3 plusieurs expériences mettant en jeu diverses techniques de traitement automatique des langues, pour en identifier les apports, mais aussi les limites. Enfin, nous discutons ces résultats dans la section 4 en rappelant les grands enjeux en cours et les perspectives très riches ouvertes par l'usage de techniques d'analyse automatique de l'information dans le cadre des sciences sociales.

ANALYSE SÉMANTIQUE AUTOMATIQUE

Cette section présente un aperçu du domaine et des techniques couramment employées aujourd'hui. Cette section comporte deux parties : la première consacrée à l'analyse factuelle (repérage d'entités nommées, de liens entre entités, etc.) tandis que la seconde partie sera consacrée à l'analyse dite subjective (analyse des sentiments, de l'opinion, etc.).

---

2. C'est en partie le rôle du TGIR (Très Grande Infrastructure de Recherche) Huma-Num (http://www.huma-num.fr/) qui peut offrir une assistance sur des outils standard. En revanche, Huma-Num ne peut pas proposer de solution originale pour chaque projet particulier, ce qui dépasserait sa mission et ses moyens.

Analyse factuelle

Avant d'en venir à l'exposé des techniques actuelles, jetons un rapide coup d'œil à l'évolution du domaine, dans la mesure où l'historique permet de bien comprendre la situation présente.

Aperçu historique

Le traitement automatique des langues (TAL) est un domaine déjà ancien, qui est apparu dès les débuts de l'informatique, après la Seconde Guerre mondiale.

La première vague de développement du TAL (1945-1965) a mis en avant la traduction automatique. Cette application est extrêmement ambitieuse dans la mesure où, idéalement, elle suppose que la machine puisse « comprendre » le texte à traduire (suivant l'adage « pour pouvoir traduire un texte, il faut d'abord le comprendre ») avant de le « reproduire » dans la langue cible (on parle aussi de « génération » pour désigner cette deuxième phase). Cette ambition dépassait largement l'état de l'art dans les années 1950 et les grands espoirs initiaux d'avancées rapides n'ont pas donné les effets escomptés. Le rapport américain ALPAC (1966) a marqué le domaine, en donnant une vision très critique des expériences menées jusque-là, ce qui a abouti à un assèchement brutal des sources de financement, au moins du côté américain (Hutchins, 2001). Ce rapport critiquait essentiellement les approches trop naïves et surtout l'absence d'analyse approfondie des textes à traduire, ce qui laissait peu d'espoir de réel progrès dans le domaine à court ou moyen terme.

La période qui a suivi (1965-1985) s'est alors, logiquement, penchée sur la question de la compréhension automatique de texte. L'intelligence artificielle occupait une grande part dans ces travaux dans la mesure où la sémantique était mise en avant, ainsi que les formalismes de représentation des connaissances (pour prendre en compte notamment le lien entre connaissances linguistiques et connaissances sur le monde, cf. Sabah, 1988). Ces recherches, nombreuses et largement financées par des subsides publics aux États-Unis, avaient abouti, dans les années 1980, à un état de l'art peu lisible. Les recherches portaient sur des types de textes différents, avaient des objectifs divers et étaient rarement évaluées. Leur déploiement en milieu opérationnel semblait très lointain et les objectifs applicatifs encore flous et incertains, ce qui était évidemment un problème pour des agences de financement ayant avant tout des objectifs appliqués (Poibeau, 2003).

Les organismes de financement américains ont alors décidé de lancer des campagnes d'évaluation afin de rendre possible la comparaison de systèmes, en développant des tâches et des jeux de données publics, communs et réutilisables. Parallèlement, des métriques automatiques étaient mises au point afin de mesurer les performances, comparer les systèmes et leur évolution dans le temps. Les premières campagnes ont porté sur la compréhension de textes (conférences MUC, *Message Understanding Conferences*, 1987-1998) et ont été suivies par d'autres sur des thèmes similaires (TREC pour la recherche d'information, DUC puis TAC pour le résumé automatique, etc.).

Les premières campagnes, qui avaient un rôle exploratoire et laissaient une grande marge de manœuvre aux participants, ont montré que la compréhension de textes était en soi une tâche floue, complexe et mal définie. Qu'est-ce que comprendre un texte ? Comment formaliser cette notion ? Quel niveau de détail faut-il prendre en compte ? Les participants et les organisateurs se mettent alors d'accord, à la fin des années 1980, sur la nécessité de limiter dans un premier temps les ambitions au repérage d'informations factuelles, locales et faciles à évaluer. Les années 1990 verront ensuite l'apparition de sous-tâches particulières, menant au développement de modules de traitement génériques et réutilisables pour différents types d'applications.

Des modules d'analyse réutilisables

Les conférences MUC ont mis en avant un certain nombre de tâches (et/ou de modules d'analyse) qui sont fréquemment reprises pour les applications visant l'analyse de contenus en langage naturel (Poibeau, 2003, 2011).

Analyse des entités nommées : les entités nommées regroupent l'ensemble des séquences faisant référence à des entités connues, comme des personnes, des lieux, des entreprises ou des organisations. Par extension, les dates et les autres expressions numériques sont fréquemment regroupées avec les entités nommées. Les termes techniques sont aussi parfois assimilés à des entités, ce qui revient alors à élargir la classe à toutes les expressions d'intérêt pour un domaine donné.

Analyse de la coréférence : une même entité peut être dénommée de façon très variée dans un même texte (par ex. *Jacques Chirac*, *le président Chirac*, *le président*, *il*…). L'analyse de la coréférence vise à reconnaître les différentes dénominations d'une même entité, ce qui a un intérêt évident pour la compréhension de textes : on peut ainsi affecter à une même entité l'ensemble

des informations qui la concernent, quelle que soit la forme sous laquelle cette entité apparaît en pratique.

Analyse des relations entre entités : cette tâche, au nom explicite, vise à identifier les relations entre entités telles qu'elles sont exprimées dans les textes. L'analyse de relations suppose une analyse correcte des prédicats, c'est-à-dire des éléments mettant en relation les différents éléments de la phrase (notamment les verbes et les noms prédicatifs) et, plus généralement, une analyse syntaxique correcte si on veut une analyse fiable et précise.

Analyse des événements : il n'y a pas de définition claire et précise de ce qu'est un événement, mais, au-delà de la simple analyse des relations, il est fréquemment nécessaire d'identifier des ensembles de plus haut niveau, rassemblant un certain nombre de relations simples dont l'agrégat est assimilé à un événement.

D'autres modules peuvent bien entendu être définis pour des besoins ou des tâches particulières. On a ainsi vu apparaître depuis quelques années une analyse plus fine des informations temporelles au sein des textes, ce qui est intéressant pour un grand nombre d'applications visant le suivi d'événements plus ou moins longs et leurs enchaînements. Les précédents modules semblent toutefois garder une plus grande généricité et être les plus communément repris au sein d'applications impliquant l'analyse de grandes masses textuelles.

Aperçu des techniques mises en œuvre

Il n'y a pas lieu de décrire ici en détail les techniques mises en œuvre. On pourra toutefois mettre en avant deux ou trois grands types d'approches :

– Dans les années 1980, la plupart des systèmes visent une analyse approfondie du contenu et, de fait, mettent en jeu des systèmes de connaissances et de représentation très fouillés. Ces systèmes sont par conséquent très coûteux à mettre en œuvre et peu portables d'un domaine à l'autre.

– Les années 1990 voient fleurir les systèmes fondés sur la technologie à nombre fini d'états (automates et/ou transducteurs à nombre fini d'états). Des résultats théoriques avaient démontré que cette technologie n'était pas suffisante pour représenter toute la complexité des langues humaines, mais, à l'inverse, plusieurs équipes montrent au cours des années 1990 que cette technologie est en fait extrêmement efficace, simple à mettre en œuvre et particulièrement appropriée pour la reconnaissance de séquences locales comme

c'est le cas dans le cadre d'une analyse sémantique locale (voir notamment Hobbs *et al.*, 1993).

– Les années 2000 voient quant à elles se généraliser le recours aux systèmes fondés sur l'apprentissage. Ces systèmes sont en théorie plus portables que les précédents, car l'expert peut se contenter d'annoter un texte et c'est ensuite la machine qui « apprend une grammaire » ou, en tout cas, des règles permettant d'annoter les textes sur la base de l'annotation manuelle[3]. Des résultats remarquables ont été obtenus ainsi, tant en qualité qu'en temps de développement (Tellier et Steedman, 2010 ; Gaussier et Yvon, 2011).

On voit aujourd'hui coexister les deux derniers types de systèmes. Le recours à l'apprentissage automatique reste un sujet de recherche et ce type de technique continue de se développer. Les entreprises commerciales ont quant à elles encore massivement recours aux systèmes à base de transducteurs à nombre fini d'états, notamment parce qu'ils offrent des qualités particulières (facilité de lecture et donc de révision par un humain ; les systèmes à base d'apprentissage artificiel sont beaucoup plus difficiles à corriger et à faire évoluer localement[4]).

Succès et difficultés

Comme on l'a vu, le traitement automatique des langues a permis des avancées majeures et est maintenant capable de fournir des modules efficaces pour traiter de grandes masses de données textuelles. Il faut toutefois souligner deux types de difficultés, ayant trait d'une part à l'analyse linguistique et de l'autre à l'ingénierie des connaissances.

En ce qui concerne l'ingénierie linguistique, on a souvent affaire à des architectures en « pipe-line » : un niveau d'analyse dépend du précédent (l'analyse sémantique repose sur l'analyse syntaxique, qui elle-même repose sur

3. Comme on verra avec les applications décrites par la suite, la réalité est toutefois plus complexe dans la mesure où les systèmes fondés sur l'apprentissage nécessitent une masse de textes annotés parfois difficile à obtenir. Pour certaines tâches simples, il peut être plus rapide de passer par un système de règles que par un système fondé sur l'apprentissage. Choisir la bonne technique pour un problème donné reste une question difficile qui nécessite souvent l'expertise d'un spécialiste du TAL.

4. Notons qu'il s'agit d'un problème difficile, car la complexité des systèmes à base de règles les rend eux aussi difficiles à faire évoluer manuellement. Globalement, la maintenabilité des systèmes de TAL reste un sujet de recherche aujourd'hui, surtout en milieu industriel (cf. note précédente).

la morphosyntaxe ou, pour prendre un exemple connexe, l'extraction d'événements dépend de l'analyse correcte de syntagmes exprimant des relations ou référant à des entités nommées). Chaque niveau d'analyse a tendance à amplifier les erreurs du niveau précédent, ce qui a évidemment une influence négative sur les performances globales. Par ailleurs, les systèmes sont peu performants dès que l'analyse dépasse les limites de la phrase, ce qui est pourtant souvent nécessaire, comme pour l'analyse des événements par exemple.

L'ingénierie des connaissances est une autre source de difficultés dès que l'on s'intéresse à des situations réelles. Par exemple, si on demande à un biologiste de reconnaître et d'annoter dans les textes des interactions entre gènes, celui-ci va avoir beaucoup de mal avec un grand nombre de cas parmi les plus importants. En effet, les cas avérés sont généralement exprimés clairement et sont relativement peu intéressants quand il s'agit d'informations connues, servant de « point de repère » (« *le texte parle ici du gène X, qui a une interaction bien établie avec le gène Y* »). Ce sont à l'inverse les cas limites qui sont précieux pour l'expert, mais problématiques pour l'analyse, car ils sont toujours exprimés avec des modalités et des prises de distance, ce qui ne permet pas de les catégoriser nettement. Un texte peut par exemple indiquer qu'il pourrait y avoir interaction entre deux gènes (clairement identifiés ou non) sans que l'auteur prenne position de façon certaine. Ce type de séquences est généralement très difficile à repérer pour le profane : on a souvent affaire à un discours très technique, modalisé, tout en nuances et qui emploie finalement peu des mots clés typiques identifiés pour la tâche. L'expert a aussi en général les plus grandes difficultés à annoter ces cas : face au linguiste, il va souvent avoir un discours complexe qui exprime une hésitation, qui « déroule » son raisonnement face au texte sans pouvoir toujours répondre de manière affirmative ou négative : « oui, il y a interaction » ou « non, ce n'est pas un cas d'interaction » (tout simplement parce que le texte ne prend pas position ainsi). Ce type de problème est aussi très présent en sciences sociales où les faits ne sont pas toujours « catégorisables » de façon claire : « annoter », c'est-à-dire « catégoriser », implique généralement de simplifier, ce qui peut s'opposer à la volonté de saisir un phénomène dans toute sa complexité.

Nous mettons ici volontairement le doigt sur les difficultés de ce type d'analyse. Il faut toutefois garder à l'esprit les grands succès obtenus depuis plusieurs années : les modules mentionnés supra montrent que des outils génériques, précis et performants existent pour plusieurs applications clés. La masse de données textuelles aujourd'hui disponible permet de concevoir

des outils fondés sur l'apprentissage automatique, qui peuvent être adaptés très rapidement à un nouveau domaine si des données représentatives sont disponibles. L'utilisation de tels outils en sciences sociales est devenue quasi indispensable pour analyser l'information sur Internet ou sur les réseaux sociaux.

Analyse subjective

L'analyse subjective, c'est-à-dire essentiellement l'analyse des sentiments et de l'opinion véhiculée dans les textes, est devenue un domaine de recherche très actif ces dernières années. On peut distinguer trois sous-tâches principales. La première sous-tâche consiste à distinguer les textes subjectifs des textes objectifs (Bethard *et al.*, 2004) ; la deuxième s'attache à classer les textes subjectifs en positifs ou négatifs (Turney, 2002) ; enfin, la troisième essaie de déterminer jusqu'à quel point les textes sont positifs ou négatifs (Wiebe *et al.*, 2001).

Plusieurs ressources ont été développées autour de l'analyse d'opinion, ou, plus largement, de tout ce qui concerne les sentiments face à un événement ou une situation donnée. On pourra citer Wordnet-Affect (Strapparava et Valitutti, 2004) ou SentiWordnet (Esuli et Sebastiani, 2006 ; Baccianella *et al.*, 2010) pour l'anglais. La première ressource est plus large que la seconde, dans la mesure où elle couvre une grande variété de sentiments, tandis que la seconde est davantage orientée vers l'analyse d'opinion. Il existe encore peu de ressources pour le français.

Des méthodes semi-automatiques destinées à compléter les ressources manuelles ont été conçues plus récemment (Vernier et Monceaux, 2007). Une étape importante pour la création de lexiques adaptés est l'identification des passages contenant des opinions (Kao et Chen, 2010) et dans ce cadre Twitter est une source importante pour l'élaboration de corpus d'opinion (Pak et Paroubek, 2010). Il faut enfin noter l'impulsion donnée par des campagnes telles que TREC Blog Opinion Task depuis 2006 (Zhang *et al.*, 2007 ; Dey et Haque, 2008) : ces campagnes ont entraîné une forte opérationnalisation du domaine et l'intégration des techniques d'analyse d'opinion dans des systèmes ouverts.

Les techniques initiales reposaient sur le repérage de mots clés porteurs de polarité (positive ou négative), ce qui permettait de calculer une valeur

d'opinion particulière au niveau de la phrase, du paragraphe ou du texte. Ces approches sont aujourd'hui considérées comme trop grossières : les articles récents font souvent appel à la théorie de l'*Appraisal* (Scherer *et al.*, 2001) qui vise à classifier les émotions en différentes catégories (peur, surprise, joie, etc.) suivant le contexte. Même si on en reste à l'analyse de l'opinion au sens étroit du terme, on voit qu'il est nécessaire de prendre en compte des aspects discursifs pour obtenir une analyse précise. Par exemple, le contraste permet d'exprimer des opinions précises (« *J'aime ce film pour son humour, mais je n'ai pas aimé les scènes de violence.* »). L'analyse doit alors prendre en compte les marqueurs de contraste (comme « *mais* ») et les éléments particuliers sur lesquels porte l'opinion (on parle de « facettes » pour désigner des éléments comme « *l'humour du film* » ou « *les scènes de violence* »). Ce type d'analyse est évidemment beaucoup plus difficile que le simple repérage de mots clés, mais les résultats sont aussi beaucoup plus précis et utilisables.

L'analyse de l'opinion met donc aujourd'hui en jeu des éléments liés au contexte linguistique, ce qui exige une analyse relativement complète des phrases visées. La négation joue aussi un rôle clé (« *j'ai aimé* » *versus* « *je n'ai pas aimé ce film* ») ; or on sait que la portée de la négation (c'est-à-dire la détermination de l'ensemble des éléments niés par la négation) est un des problèmes les plus difficiles de la linguistique. Il est donc intéressant de constater que l'analyse d'opinion, qui a pu apparaître dans un premier temps comme un domaine relativement superficiel et appliqué (fondé initialement sur le simple repérage de mots clés avec polarité), finit par mettre le doigt sur certains des aspects les plus difficiles de l'analyse linguistique, de l'analyse du discours à celle des modalités en passant par la négation.

QUELQUES EXPÉRIENCES RÉCENTES

Nous présentons dans cette section quelques réalisations récentes utilisant les technologies présentées ci-dessus. À travers ces différents exemples, on verra que la part liée au TAL est variable et que les techniques utilisées évoluent elles aussi en fonction du contexte. Les difficultés récurrentes concernent l'adaptabilité des systèmes, le temps nécessaire pour développer des ressources pour un nouveau domaine et la disponibilité ou non d'exemples en quantité suffisante (c'est-à-dire des données annotées représentatives du problème visé). La définition des informations recherchées et la qualité des résultats obtenus sont également des questions régulièrement posées.

Nous présentons donc, dans l'ordre :

– Une expérience visant à reconstituer la généalogie des citations lors de la campagne présidentielle américaine de 2008. Le traitement de ce type d'information est important, mais nécessite finalement relativement peu de techniques issues du TAL (Omodei *et al.*, 2012).

– Une application d'extraction d'information à partir d'un fil d'agence de presse (Poibeau, 2002). Il s'agit ici de reconnaître des informations précises, évolutives, sur des thèmes variés. L'adaptabilité d'un tel système est un problème majeur.

– Un système de résumé automatique de textes d'opinion (Poibeau, 2011). Cette application pose des questions de recherche redoutables, car elle nécessite de traiter conjointement information factuelle et subjective.

– Enfin, nous présentons une expérience préliminaire visant directement l'application de techniques issues du TAL pour les sciences sociales (Bourreau et Poibeau, 2014).

Il y a donc un continuum entre des applications analysant des textes, mais s'intéressant peu au contenu jusqu'à des applications s'attachant à une analyse fine de leur contenu objectif et subjectif.

Analyse de la diffusion de l'information : l'exemple de la campagne présidentielle américaine de 2008

La première application concerne l'analyse d'un jeu de données concernant des reprises de citations lors de la campagne présidentielle américaine de 2008.

Présentation de l'application

Lors d'une campagne électorale, certains thèmes émergent et s'imposent dans le débat, tandis que d'autres passent rapidement au second plan et ne sont plus traités par la suite. De même, certains thèmes suscitent de multiples réactions, qui sont elles-mêmes reprises, commentées, discutées. Le suivi de ces citations est potentiellement source de connaissances, mais une analyse manuelle est impossible dans la mesure où il faut détecter les citations, identifier leurs sources, leur déformation, etc. dans des corpus immenses sur Internet. L'analyse automatique peut fournir des pistes précieuses dans ce contexte.

Aperçu technique

L'analyse des citations implique, au moins dans un premier temps, très peu de connaissances d'ordre linguistique. À l'image de l'analyse des manuscrits anciens, la filiation des citations est reconstituée en restant au plus proche du texte, qui doit de plus pouvoir être daté. Chaque modification est repérée, ce qui permet de reconstituer de manière relativement fiable la généalogie des citations (à l'image de l'analyse philologique, qui part de l'étude des variations entre manuscrits pour reconstituer leur généalogie, les textes étant produits par copies successives et donc reprenant les modifications précédentes d'une version à l'autre).

Les citations peuvent ensuite être regroupées en différentes familles reflétant leur proximité relative. Une modélisation fine du phénomène permet en outre de pondérer les différents facteurs observés et de montrer qu'ils permettent effectivement d'expliquer les observations faites à partir des données réelles.

Discussion

Cette analyse permet de reconstituer le circuit de circulation des citations : comment elles sont reprises, transformées, modifiées. L'analyse révèle que les citations les plus populaires sont aussi les plus stables. Certains thèmes sont repris, figés et acquièrent progressivement le statut de slogan. Il est difficile d'expliquer ce phénomène, c'est-à-dire de mettre au jour les éléments qui font qu'un thème ou qu'une citation se fige et acquiert un statut spécial. Il n'empêche que les outils automatiques permettent une détection rapide et efficace, ils peuvent en outre donner des éléments chiffrés pour fournir une base solide à tout essai d'interprétation.

L'analyse éclaire en outre le fonctionnement de ces phénomènes de reprise et de citation : quand une citation est longue, elle est souvent reprise telle quelle, alors que les citations plus courtes sont davantage sujettes à modification. On peut imaginer que les citations courtes (jusqu'à huit mots) sont citées de mémoire sans vérification et, du coup, subissent davantage de modifications. À l'inverse, les citations longues sont plus difficiles à retenir par cœur et sont donc reprises par copier-coller, ce qui ne peut donner lieu à modification involontaire.

Enfin, une étude plus fine du contenu des citations serait nécessaire pour mieux voir les thèmes repris et ceux laissés de côté. Une bonne connaissance de la structure du réseau social ainsi que de l'orientation politique des auteurs

de citations (*qui reprend quoi* ?) constituerait également une source d'enseignements précieux. Il faut toutefois noter que ce type d'applications n'est pas sans poser de multiples questions éthiques, car il est alors possible de suivre l'opinion et potentiellement de la manipuler (nous avons travaillé ici avec un jeu de données publiques, mais de nombreux acteurs du Web maîtrisent ce type de techniques qu'ils peuvent appliquer aux données qu'ils stockent, publiques ou privées).

Suivi de thèmes dans un corpus d'agence de presse

L'analyse d'un fil de presse pose des problèmes classiques en matière d'analyse d'informations textuelles : les systèmes d'extraction d'information sont généralement spécialisés et capables d'analyser un domaine particulier. À l'inverse, un fil d'agence de presse porte sur des thèmes variés : il faut donc concevoir un système rapidement adaptable en fonction des différents thèmes abordés.

Présentation de l'application

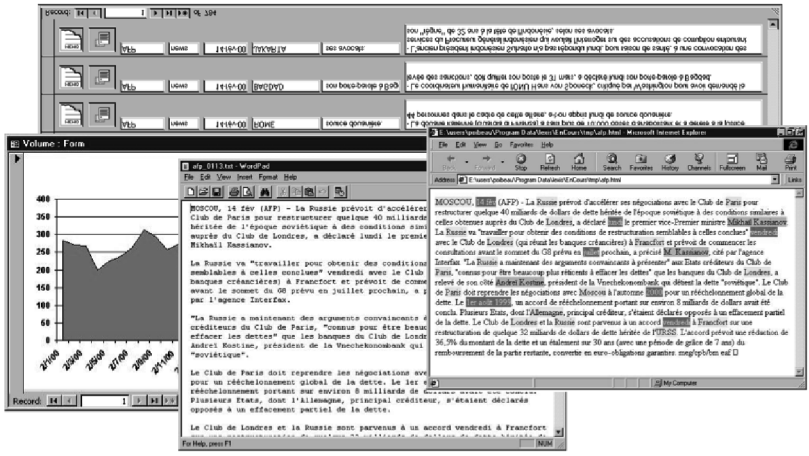
L'application porte sur l'analyse en direct d'un fil d'agence de presse en français (figure 1). Le système visé se compose donc de deux modules principaux : un premier module cherche à identifier les thèmes abordés dans les dépêches, puis une application classique d'extraction d'information remplit un formulaire spécifique si un thème connu a été identifié.

Le principal enjeu d'un tel système est évidemment sa maintenance et son évolutivité. Pour être pertinente, l'information extraite doit être précise et adaptée en fonction des événements en cours. Il est donc nécessaire que l'analyste puisse développer de nouvelles ressources très rapidement en fonction de l'actualité.

Aperçu technique

Si l'apprentissage artificiel est aujourd'hui massivement utilisé pour les applications d'extraction d'information, les systèmes opérationnels restent encore en majorité fondés sur des techniques symboliques, essentiellement des transducteurs à nombre fini d'états. Ceux-ci présentent plusieurs avantages, dans la mesure où ils permettent d'utiliser des ressources linguistiques comme des dictionnaires et des grammaires explicites, plus faciles à faire évoluer que des ressources acquises automatiquement par des techniques d'apprentissage artificiel. L'apprentissage automatique requiert aussi de disposer de masses de

Figure 1. Analyse dynamique de dépêche. Le texte est d'abord enrichi puis les informations essentielles sont extraites et stockées dans une base de données.



données importantes, le plus souvent annotées, qui ne sont pas toujours disponibles, surtout dans le cadre d'une application visant l'analyse de fil d'actualité (portant donc sur des thèmes changeants et peu prédictibles).

Les techniques d'apprentissage artificiel ne sont toutefois pas complètement laissées de côté. Certains thèmes sont plus stables que d'autres, mais évoluent néanmoins. Les utilisateurs peuvent alors reclasser certaines dépêches, ou montrer leur intérêt pour de nouvelles dépêches non classées ou mal classées jusque-là. L'apprentissage interactif et/ou incrémental permet alors de faire évoluer le système dynamiquement (les nouvelles données permettant de compléter automatiquement les connaissances du système) à partir des « retours » des utilisateurs.

Discussion

Le suivi de thème au sein d'un fil d'agence de presse semble à première vue constituer une application typique de l'extraction d'information. Ce qui est cherché est généralement une information factuelle et vérifiable. Dans les faits, il en va assez différemment. Un grand nombre de dépêches, dans le domaine politique notamment, mais pas seulement, concerne des réactions, des opinions sur des faits récents, voire des opinions sur des opinions. Ce type d'information n'est pas pris en compte ici, mais il est important de noter le continuum entre information factuelle et information plus subjective.

L'adaptabilité des systèmes est le problème clé de toute application visant l'analyse du langage naturel : la mise au point de systèmes adaptatifs reste un domaine de recherche à part entière. La découverte dynamique de thèmes au sein d'un corpus, voire le développement d'applications définissant les formulaires d'extraction « à la volée », commence à être explorée du point de vue de la recherche, mais on est encore loin d'avoir de tels systèmes opérationnels (Wei *et al.*, 2013).

Résumé automatique de textes d'opinion

Présentation de l'application

Le résumé automatique de textes est un domaine de recherche qui a récemment connu un essor important : cette technologie a un intérêt particulier quand l'utilisateur doit faire face à une masse de documents importante dont il faut prendre connaissance en temps limité. Le résumé automatique de textes permet de repérer les documents et les faits importants, de mettre ceux-ci en avant et de jouer le rôle de « filtre » pour accéder rapidement à l'information pertinente (autrement dit, hiérarchiser l'information selon son importance).

Les campagnes TAC (*Text Analysis Conference*) organisées chaque année depuis 2008 par le NIST (*National Institute of Standards and Technology*, le NIST est un organisme public américain) ont proposé plusieurs évaluations autour de la problématique du résumé, de l'analyse d'opinion et de la reconnaissance d'événements au sein de corpus variés. Nous rapportons ici une expérience qui a eu lieu en 2008, lors d'une campagne sur le résumé automatique de textes rapportant des avis et des opinions sur différents thèmes.

Approche

L'approche traditionnelle en résumé automatique vise à repérer les phrases importantes dans un premier temps, puis à essayer d'éliminer les phrases redondantes dans un deuxième temps. Cette approche, bien qu'elle soit très répandue, ne semble pas idéale : il serait plus logique de classer d'abord les phrases suivant la nature de l'information rapportée, puis de choisir une ou plusieurs phrases représentatives pour produire le résumé, plutôt que de faire les choses en ordre inverse.

L'approche que nous avons retenue pour le résumé consiste à enrichir les documents avec différents types d'annotations, afin de normaliser les phrases

et identifier celles qui rapportent le même type d'information (les annotations fonctionnant comme des catégories génériques pouvant correspondre à différentes formulations linguistiques). On sait en effet qu'une des principales difficultés de l'analyse automatique de textes se situe dans la reconnaissance d'informations similaires, dans la mesure où une même information peut être exprimée de façon infiniment variée du fait de la plasticité de la langue. Cette plasticité est liée à des phénomènes de nature variée : choix de mots différents pour exprimer une même notion (emploi de synonymes, hyponymes, etc.), emploi de structures différentes (phrases simples, complexes, à l'actif ou au passif, etc.) mais quasi équivalentes sur le plan sémantique (paraphrases), etc. La combinatoire et la variété de ces sources de variation font que l'on ne peut pas lister *a priori* l'ensemble des façons d'exprimer une notion. Les systèmes fonctionnent alors à rebours : il s'agit d'inférer, à partir du repérage d'un ensemble d'éléments proches ou similaires, que deux phrases expriment la même notion (le même fait, le même événement).

L'analyse d'opinion pose des problèmes en partie similaires à la reconnaissance d'événements : la façon d'exprimer une opinion peut être très variée, le vocabulaire employé change fortement d'un domaine à l'autre et une analyse précise demande fréquemment le recours à un contexte large. Du fait de ces difficultés, nous avons rapidement abandonné l'idée de partir d'un lexique de mots avec polarité (lexique où des mots comme « *bon* » ou « *joie* » sont associés à une valeur positive et des mots comme « *mauvais* » ou « *triste* » avec une valeur négative) : ces lexiques sont très incomplets et ne sont pas toujours justes face à un corpus particulier. Il semble plus prometteur de partir de techniques d'apprentissage artificiel pour créer dynamiquement des lexiques adaptés au domaine visé.

La production du résumé intervient alors sur la base des informations précédemment rassemblées. Le calcul de la proximité entre phrases permet de regrouper les phrases en classes d'équivalence (comme il s'agit de résumé multi-documents, plusieurs phrases de différents documents sources peuvent rapporter la même information, surtout s'il s'agit d'une information essentielle ; le nombre de phrases en situation d'équivalence sémantique est d'ailleurs un indice majeur pour identifier les informations essentielles dans un texte).

Le système de résumé identifie alors une phrase centrale pour chacun des principaux regroupements et ordonne ces phrases de façon à former un tout si possible cohérent. Dans le cas de résumés d'opinion, le choix est

généralement de regrouper les phrases suivant les opinions exprimées (les opinions positives d'abord, puis les opinions négatives), mais d'autres choix sont possibles. Différents mécanismes vérifient par ailleurs que le résumé produit est conforme à la longueur visée si celle-ci a été définie à l'avance (il peut s'agir d'une longueur absolue ou relative à la taille du corpus de départ). Ces résumés peuvent ensuite être évalués suivant différents critères.

Résultats et discussion

Il existe principalement deux manières d'évaluer un résumé produit automatiquement. La première façon de procéder est classique : un expert du domaine considéré lit le résumé produit et donne une ou plusieurs notes censées illustrer la qualité finale et/ou différents critères supposés pertinents pour la tâche. La deuxième méthode consiste à élaborer un algorithme permettant de juger automatiquement un résumé. Il va de soi que cette deuxième façon de faire est très complexe, d'autant qu'on ne sait pas formaliser la notion de « bon résumé ». Même entre experts, les notes et les évaluations varient de façon notable, il est donc logique qu'il en aille de même pour un système automatique.

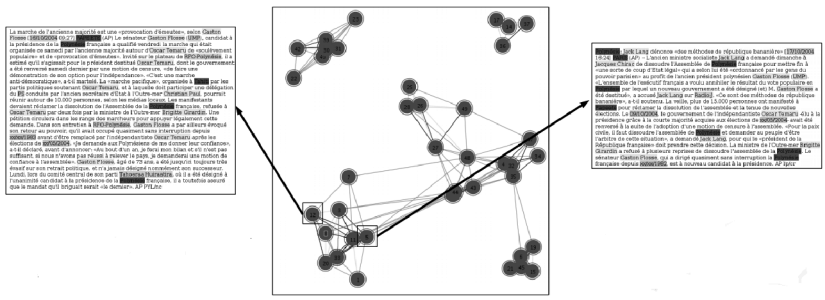
Le système que nous avons développé a obtenu de bons scores relatifs, se glissant en tête des systèmes évalués. Les autres tests ont toutefois révélé des résultats globalement moyens, c'est-à-dire que le système se comporte bien par rapport aux autres systèmes équivalents, mais que les résultats produits restent médiocres quant à la qualité des textes et à leur lisibilité.

Ces résultats ne sont guère surprenants et doivent être interprétés à l'aune de l'état de l'art : si l'extraction d'informations factuelles et locales est une technologie correctement maîtrisée, la production de textes longs, réalistes et cohérents, reste un problème difficile. De même, identifier avec précision l'importance d'une information par rapport à une autre est quasiment au-delà de l'état de l'art (et, là encore, il s'agit d'un domaine très subjectif où même des évaluateurs humains ont des avis relativement peu stables).

Il n'empêche, cette expérience montre qu'il est malgré tout possible d'analyser de grandes masses de données, d'en extraire des informations pertinentes, et de les mettre en forme, même si l'on vient de voir que cette dernière tâche est à la limite de l'état de l'art. De manière plus intéressante, on pourra remarquer que, dans l'approche que nous avons proposée, la détection de la redondance passe par le regroupement de phrases évoquant des idées proches :

l'utilisateur peut avoir accès à ce résultat intermédiaire sans se focaliser spécialement sur le résultat final (le résumé produit). Les résultats intermédiaires sont disponibles et utilisables grâce à des représentations variées (nous avons par exemple imaginé des représentations graphiques permettant à l'utilisateur de voir les regroupements effectués et de naviguer de manière interactive dans les cartes ainsi produites, en partant du nom des principaux acteurs concernés ou d'autres données intéressantes dans le cadre de sa recherche), cf. Bossard et Poibeau, 2008, ainsi que la figure 2.

Figure 2. Regroupements de documents en classes événementielles. Chaque point renvoie à un document qui peut être visualisé à la demande. La même représentation est possible au niveau des phrases.



Les résultats produits sont donc multiples et les étapes intermédiaires de l'analyse sont probablement aussi intéressantes que les résultats finaux, à savoir les résumés en langage naturel produits automatiquement.

Une application directement liée aux sciences sociales : la campagne PoliInformatics 2014

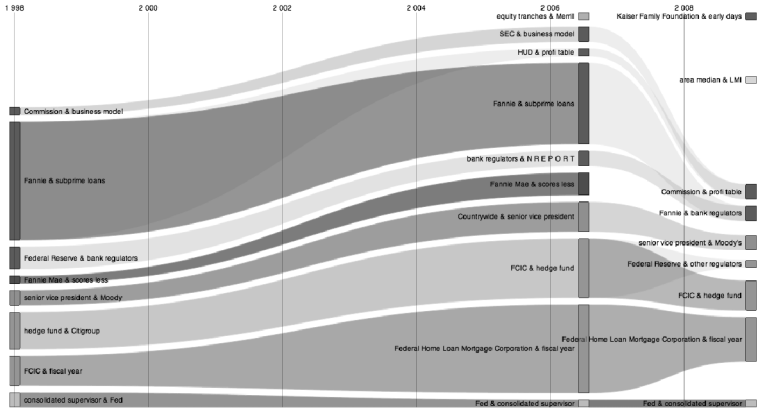
On assiste depuis quelque temps à une multiplication des projets impliquant le TAL au service des sciences sociales. Pour prendre un exemple récent, une initiative d'origine américaine appelée PoliInformatics (http://poliinformatics.org/) a ainsi proposé d'examiner en quoi les techniques de TAL peuvent apporter une aide aux chercheurs en droit et en sciences politiques face à des événements complexes ayant produit une grande masse de données, essentiellement sous forme textuelle. Une campagne d'évaluation exploratoire a été menée en 2014 : les données fournies concernaient la réponse du gouvernement et des autorités américaines à la crise financière de 2008-2009.

Ce type d'initiative est intéressant à plusieurs titres. D'une part, la tâche n'est pas complètement spécifiée à l'avance : dans le cadre de PoliInformatics, les organisateurs avaient juste imaginé les questions suivantes : "*Who was the financial crisis?*" et "*What was the financial crisis?*". Autrement dit, le but consistait à identifier et caractériser les principaux acteurs de la crise américaine et, plus largement, à fournir des éléments d'information sur les éléments saillants de la crise financière (on peut ainsi imaginer que les experts sont intéressés par les causes de la crise, mais aussi et peut-être plus subtilement, par les prises de position des différents acteurs, leurs explications et les contradictions qui peuvent apparaître entre les acteurs).

Nous avons pour notre part fait le choix, dans un premier temps, de juste extraire certaines informations essentielles (entités nommées et liens entre entités) afin de permettre à l'utilisateur de voir les liens entre les acteurs du domaine et de naviguer au sein du réseau ainsi obtenu. Une autre représentation utile est l'évolution des thèmes au cours du temps (figure 3).

À ce propos, il faut rappeler que dans ce type de tâche, les outils automatiques ne peuvent fournir la « bonne réponse » dans la mesure où les événements sont complexes, et la perception que l'on peut en avoir dépend largement du point de vue et de la position des acteurs par rapport aux faits évoqués. Les outils automatiques sont donc utiles, mais ils ne peuvent faire qu'une partie

Figure 3. Identification de thèmes et visualisation de leur évolution au cours du temps dans une partie du corpus PoliInformatics. Cette visualisation est obtenue grâce à la plate-forme Cortext de l'IFRIS.



du travail et en aucun cas se substituer à l'expert ou à l'analyste. Grâce aux techniques évoquées précédemment, il est en effet possible d'identifier des arguments, de regrouper les différentes positions sur un même fait de façon rapide et interactive afin de faciliter l'analyse. Ces outils peuvent être particulièrement pertinents pour identifier les différents points de vue sur un même événement et caractériser ces différences (au niveau du lexique, des modalités ou des tournures employées par exemple).

L'interprétation des données reste en revanche entièrement à la charge de l'expert. On peut ici se rappeler ce que l'on a observé pour le résumé automatique de texte : les outils ne sont à l'heure actuelle pas entièrement fiables pour identifier l'importance d'une information relativement à d'autres ou pour vérifier la validité d'une source de données. En revanche, les outils automatiques sont indispensables pour parcourir rapidement de très grandes masses de documents qui seraient sinon très difficiles à analyser par des experts. Les outils peuvent aussi contribuer à donner une vue globale sur un corpus par exemple, et différents moyens d'accès à l'information à travers des cartes et des outils de navigation interactifs.

DISCUSSION

Ces expériences nous semblent intéressantes à plusieurs titres : elles montrent qu'il est désormais possible d'analyser automatiquement des données textuelles de manière relativement fine. Cette analyse automatique ne se substitue pas à une analyse humaine, mais peut l'assister efficacement pour manipuler automatiquement de grandes masses de documents, qui dépassent aujourd'hui fréquemment les capacités de lecture du chercheur ou de l'analyste.

Les outils automatiques : des assistants précieux, mais à manier avec précaution

On a montré dans les pages qui précèdent les succès importants des outils de traitement automatique des langues. Identifier les noms de personnes, les termes essentiels, les entités ainsi que les relations entre entités sont des tâches bien maîtrisées avec des taux de succès satisfaisants pour la plupart des utilisations pratiques. De plus en plus de milieux professionnels ont recours à ce type d'outils pour manipuler une littérature sans cesse plus volumineuse.

Il faut cependant noter que les outils automatiques ne se substituent pas à l'expertise humaine. La machine reste incapable de faire une évaluation de l'importance et de la pertinence de l'information recherchée. Cette évaluation exige de faire des inférences à partir des données et suppose généralement une connaissance de l'état du monde ou au moins du contexte du problème traité. Or les systèmes se fondent sur des algorithmes génériques qui restent à la surface des choses. Il faut donc rester prudent dès qu'il s'agit d'analyser et d'évaluer les informations extraites automatiquement : cette tâche reste encore l'apanage de l'expert, et c'est une bonne chose !

Il en va de même pour le domaine aujourd'hui florissant de l'analyse d'opinion. Les expériences ont montré que la caractérisation fine de l'opinion (intensité, gradation de l'opinion) donnait des résultats très mitigés. D'un autre côté, les résultats obtenus par des évaluateurs humains restent eux-mêmes très variables. Ces notions sont donc probablement trop floues et trop subjectives pour faire l'objet d'une évaluation précise (en revanche, une analyse en termes de polarité peut obtenir des résultats très fiables et intéressants).

L'analyse d'opinion est plus intéressante, nous semble-t-il, quand on s'intéresse à ce qu'elle met en jeu, à savoir un ensemble de données assez largement laissées de côté jusqu'à récemment, la priorité ayant longtemps été donnée aux informations dites factuelles. L'avènement du Web participatif, à travers notamment les forums et les avis de consommateurs, a eu un poids considérable pour le renforcement des recherches en ce domaine.

La place de l'analyste face à la masse des données

Au-delà de ces remarques sur des applications particulières, on peut s'interroger sur le rôle des données massives pour le TAL. On sait que ces données ont profondément changé le domaine : l'apprentissage automatique est aujourd'hui la technologie la plus efficace dans bien des cas (Gaussier et Yvon, 2011 ; Tellier et Steedman, 2010). Même la traduction automatique est aujourd'hui fondée sur des systèmes entièrement statistiques. Contrairement à ce qui a longtemps prévalu, nul besoin de modéliser le contenu, c'est-à-dire la sémantique d'un texte pour le traduire : les données bilingues disponibles sur le Web sont plus efficaces pour fournir des bribes de traduction qui sont ensuite assemblées automatiquement. Plus généralement, une approche brute par ordinateur est souvent plus efficace que le travail de dizaines de linguistes essayant de développer à la main des dictionnaires et des règles d'analyse

adaptées (Poibeau, 2014). Les systèmes statistiques sont aussi les plus efficaces pour la recherche d'information, la classification automatique voire le résumé, comme on l'a vu ci-dessus.

L'apprentissage automatique souffre toutefois de limitations sérieuses dans certains contextes : pour pouvoir « entraîner » un système, il faut disposer de données en grandes quantités ; ces données doivent de plus être annotées afin de « guider » le système et lui indiquer les éléments pertinents. Du coup, ces techniques ne sont pas toujours faciles à utiliser, notamment quand le système doit être rapidement adapté sans que l'on dispose de données annotées (comme dans le cas de l'application d'extraction d'information décrite dans la section précédente).

Il est dès lors utile de distinguer différents cas d'usage : les systèmes statistiques sont souvent efficaces pour découvrir une information nouvelle, par exemple en détectant des co-occurrences de mots clés non observées jusque-là. Cette technique peut être utilisée pour reconstituer l'évolution d'un domaine scientifique (Chavalarias et Cointet, 2013) ou découvrir dynamiquement des thèmes intéressants dans un contexte donné (Eichler *et al*., 2008). Les systèmes à base de règles sont eux très efficaces quand une information de surface peut être trouvée rapidement et simplement par quelques règles locales. Ces règles permettent d'extraire certains éléments centraux qui peuvent à leur tour servir de base à l'acquisition dynamique de nouvelles connaissances par apprentissage incrémental et/ou interactif (on parle alors d'amorçage ; les Anglo-Saxons parlent quant à eux de « *seed* » pour désigner les « graines » qui permettent de développer un système plus performant par la suite, cf. Xu *et al.*, 2007). Les systèmes sont dits « hybrides » quand ils mélangent des règles et des approches numériques, c'est-à-dire du symbolique et du statistique : c'est sans doute l'approche la plus prometteuse pour l'avenir, même si elle reste difficile à mettre en œuvre en pratique (que faut-il définir sous forme de règle ? Quel type de connaissance peut assurer le meilleur « amorçage » pour la suite ?).

Le plus important consiste sûrement à déterminer la façon d'intégrer l'analyste dans la boucle d'analyse, la stratégie à définir dépendant d'ailleurs étroitement du cadre visé (Pierce et Cardie, 2001). Les outils automatiques sont efficaces pour identifier des thèmes, donner une idée du contenu d'un corpus documentaire trop grand pour être étudié manuellement. Les outils peuvent aussi permettre de « sonder » le contenu, c'est-à-dire estimer le contenu

probable en lançant des requêtes *a priori* pertinentes. C'est ensuite à l'analyste de définir l'information à extraire et, surtout, de décrire comment celle-ci peut être extraite. Il n'y a donc pas d'opposition entre le qualitatif et le quantitatif, de même qu'il n'y a dans le fond pas de véritable opposition entre systèmes à base de règles et systèmes statistiques, ou entre précision de l'analyse et grand corpus. C'est dans la façon de définir les problèmes, de choisir les bonnes techniques et surtout de faire collaborer le tout en un ensemble harmonieux que se situent les enjeux pour les sciences sociales.

CONCLUSION

Cet article a permis d'offrir un aperçu des outils de TAL aujourd'hui disponibles pour l'analyse de larges ensembles de textes en sciences sociales. Ces outils sont de deux types : ceux qui s'intéressent à l'information factuelle exprimée (identification des entités, des relations entre entités et des événements) et ceux qui portent sur l'information subjective (analyse de l'opinion et des sentiments). Ces outils sont aujourd'hui matures pour une aide à l'analyse de corpus, l'extraction d'informations essentielles et la navigation dans de grands ensembles de données. En revanche, la mise en évidence des faits importants, des relations entre les faits et surtout leur interprétation échappe encore grandement à la machine, même si ces différents points font l'objet de recherches actives aujourd'hui.

Les enjeux sont donc importants et largement en phase avec les besoins en sciences sociales et, plus généralement, ceux de la société de l'information. On voit par exemple se développer le *fact checking* dans la presse, c'est-à-dire la vérification d'affirmations par des experts ou des hommes politiques. Cette vérification qui demande un important travail manuel pourrait être largement assistée par des techniques automatiques même s'il faut rester prudent sur le degré d'automatisation possible (Vlachos et Riedel, 2014). Des recherches similaires sont en cours dans de multiples domaines connexes, afin par exemple d'identifier des schémas récurrents de comportements socio-économiques (Lampos *et al.*, 2014) ou l'analyse du marché de l'art, pour prendre quelques exemples variés (Al Tantawy *et al*., 2014).

Il s'agit donc d'un secteur de recherche très florissant avec des enjeux majeurs. Il s'agit aussi et surtout d'un domaine clé pour l'analyse quali-quantitative dans la mesure où la masse de textes est là, mais nécessite des traitements précis pour avoir accès à une analyse sémantique fine.

RÉFÉRENCES

ACHANANUPARP, P., HU, X., SHEN, X. (2008), *The Evaluation of Sentence Similarity Measures.* Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery, Turin (Italy), Springer-Verlag.

ALTANTAWY, M., RULE, A., RAMBOW, O., WANG, Z., BASU, R. (2014), *Using Simple NLP Tools to Trace the Globalization of the Art World.* Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Baltimore, USA.

BACCIANELLA, S., ESULI, A. SEBASTIANI, F. (2010), *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*. Proceedings of the conference on Language Resources and Evaluation (LREC'10), Malte.

BETHARD, S., YU, H., THORNTON, A., HATZIVASSILOGLOU, V., JURAFSKY, D. (2004), « Automatic extraction of opinion propositions and their holders », *Working Notes of the AAAI Spring Symposium on Exploring Attitude and Aect in Text: Theories and Applications*. Stanford.

BOSSARD, A., POIBEAU, T. (2008), *Regroupement automatique de documents en classes événementielles.* Actes de la conférence Traitement Automatique de Langage Naturel (TALN 2008), Avignon.

BOSSARD, A., GENEREUX, M., POIBEAU, T. (2010), « Résumé automatique de textes d'opinion », *Traitement Automatique des Langues*, 51/3, pp. 47-73.

BOURREAU, P., POIBEAU, T. (2014), « Mapping the Economic Crisis: Some Preliminary Investigations », *Technical report on the ACL 2014 PoliInformatics Unshared task*.arXiv:1406.4211.

CHAVALARIAS, D., COINTET, J.-P. (2013), « Phylomemetic Patterns in Science Evolution. The Rise and Fall of Scientific Fields », *PLoS One* 8(2): e54847. Doi:10.1371/journal.pone.0054847.

DEY, L., HAQUE, M. (2008), « Opinion mining from noisy text data », *AND 08: Proceedings of the secondworkshop on Analytics for noisy unstructured text data*. New York.

EICHLER, K., HEMSEN, H., LÖCKELT, M., NEUMANN, G., REITHINGER, N. (2008), « Interactive Dynamic Information Extraction », *Annual German Conference on Artificial Intelligence*, Kaiserslautern, Germany.

ESULI, A., SEBASTIANI, F. (2006), *SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining.* Proceedings of the conference on Language Resources and Evaluation (LREC'06), Gênes.

GAUSSIER, É., YVON, F. (2011), *Modèles statistiques pour l'accès à l'information textuelle*. Paris, Lavoisier, coll. « Recherche d'information et web ».

HOBBS, J. R., APPELT, D. E., BEAR, J., ISRAEL, D., KAMEYAMA, M., TYSON, M. (1993), « FASTUS: A System for Extracting Information from Text », *Proceedings of the Human Language Technology Conference*, Princeton, New Jersey, pp. 133-137.

HUTCHINS, J. (2001), « Machine translation over fifty years », *Histoire, épistémologie, langage*, vol. 23 (1), pp. 7-31.

KAO, H.-A., CHEN, H.-H. (2010), *Comment Extraction from Blog Posts and Its Applications to Opinion Mining.* Proceedings of the conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, 2010.

LAMPOS, V., PREOTIUC-PIETRO, D., SAMANGOOEI, S., GELLING, D., COHN, T. (2014), *Extracting Socioeconomic Patterns from the News: Modelling Text and Outlet Importance Jointly*. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, USA.

OMODEI E., POIBEAU, T., COINTET, J.-P. (2012), *Multi-Level Modeling of Quotation Families Morphogenesis.* Proceedings of the 2012 ASE/IEEE International Conference on Social Computing, Amsterdam.

PAK, A., PAROUBEK, P. (2010), *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*. Proceedings of the conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta.

PIERCE, D., CARDIE, C. (2001), *User-Oriented Machine Learning Strategies for Information Extraction: Putting the Human Back in the Loop. Loop*. Proceedings of the IJCAI Workshop on Adaptive Text Extraction and Mining (ATEM), Seattle, USA.

POIBEAU, T. (2002), *Extraction d'information à base de connaissances hybrides.* Thèse, Université Paris-Nord.

POIBEAU, T. (2003), *Du texte brut au web sémantique*. Paris, Lavoisier.

POIBEAU, T. (2011), *Traitement automatique du contenu textuel*. Paris, Lavoisier.

POIBEAU, T. (2014), « La linguistique est-elle soluble dans la statistique ? », *Revue Sciences/Lettres*, n° 2. URL : http://rsl.revues.org/402 ; DOI : 10.4000/rsl.402.

SCHERER K., SCHORR, A., JOHNSTONE, T. (dir.) (2001), *Appraisal processes in emotion: Theory, methods, research.* New York, Oxford University Press.

SABAH, G. (1988), *L'intelligence artificielle et le langage.* I. *Représentation des connaissances*, Paris, Hermès, 358 p.

STRAPPARAVA, C., VALITUTTI, A. (2004), « WordNet-Affect: an affective extension of WordNet », *Proceedings of the conference on Language Resources and Evaluation* (LREC'14), Lisbonne, pp. 1083-1086.

TELLIER, I., STEEDMAN, M. (2010), « Apprentissage automatique pour le TAL », *Traitement Automatique des Langues* (TAL), 50/3.

TURNEY, P. D. (2002), *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphie.

VENTURINI, T., GUIDO, D. (2012), « Once Upon a Text: an ANT Tale in Text Analysis », *Sociologica* (Italian Journal of Sociology Online), n° 3.

VERNIER, M., MONCEAUX, L. (2007), « Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques », *Traitement Automatique des Langues*, 2007.

VLACHOS, A., RIEDEL, S. (2014), *Fact Checking: Task definition and data set construction.* Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science. Baltimore, USA.

WANG W., BESANÇON, R., FERRET, O., GRAU, B. (2013), « Regroupement sémantique de relations pour l'extraction d'information non supervisée », *20e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, Les Sables d'Olonne, pp. 353-366.

WIEBE, J., WILSON, T., BELL, M. (2001), *Identifying Collocations for Recognizing Opinions. Proceedings of the ACL 2001Workshop on Collocation*. Toulouse.

XU, F., USZKOREIT, H., LI, H. (2007), « A Seed-driven Bottom-up Machine Learning Framework for Extracting Relations of Various Complexity », *Proceedings of ACL 2007, 45th Annual Meeting of the Association for Computational Linguistics*, Prague, pp. 584-591.

ZHANG, W., YU, S., MENG, W. (2007), *Opinion retrieval from blogs*. CIKM '07: Proceedings of the sixteenth ACM Conference on Information and Knowledge Management. New York.