Le traitement automatique des langues pour les sciences sociales : quelques éléments de réflexion à partir d'expériences récentes

Thierry Poibeau

Le traitement automatique des langues pour les sciences sociales : quelques éléments de réflexion à partir d'expériences récentes

Thierry Poibeau
Laboratoire LATTICE (UMR8094)
CNRS & Ecole normale supérieure & Université Paris 3 Sorbonne Nouvelle
1, rue Maurice Arnoux
92120 Montrouge France
thierry.poibeau@ens.fr

1 Introduction

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 facile d'accès. Mais, loin de l'Eldorado promit 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 triés, filtrés, organisés ; elles reflètent des points de vue particuliers qui ne sont pas obligatoirement ceux visés par le chercheur en science 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 rapide 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 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 les « bons » traitements aux chercheurs en sciences sociales en utilisant une combinaison d'outils existants, sans qu'il y ait forcément une dimension fondamentalement nouvelle dans les traitements effectués[2].

Nous laisserons là les considérations sociologiques liées à l'état de la science 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 l'analyse des opinions et des sentiments. Nous détaillons ensuite dans la section 3 une expérience autour du résumé automatique de textes d'opinions, pour en identifier les enjeux principaux. Cette expérience est intéressante en ce qu'elle implique à la fois l'identification d'informations factuelles et d'informations plus subjectives. Nous montrons à chaque fois les apports mais aussi les limites des techniques existantes. Nous concluons 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.

2 Analyse sémantique automatique

Cette section présente un aperçu rapide 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.).

2.1 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.

2.1.1 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. Les années 1980-1990 quant à elles ont été marquées par la création et le développement de conférences d'évaluation, en

[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

[2] C'est en partie le rôle du TGIR (Très Grande Infrastructure de Recherche) Huma-Num (http://www.huma-num.fr/) mais son rôle est limité à une assistance standard, en pratique très éloignée des besoins très divers exprimés par les utilisateurs en sciences sociales. Surtout, Huma-Num n'a pas les moyens humains d'assister chaque projet en particulier, ce qui n'est d'ailleurs pas son rôle.

premier lieu dans le monde anglo-saxon, mais pour comprendre la situation au milieu des années 1980, il est utile d'examiner brièvement l'histoire du domaine.

La première vague de développement du TAL (1945-1965) avait en effet vu le développement rapide puis l'échec encore plus soudain de la traduction automatique, supposée trop complexe car nécessitant au préalable une analyse sémantique fouillée des textes à traduire (Hutchins, 2001). La période qui a suivi (1965-1985) a alors vu l'éclosion de nombreuses théories et la mise au point de systèmes de compréhension de textes. L'intelligence artificielle occupe une grande part dans ces travaux dans la mesure où la sémantique est 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. Sabbah, 1988). Ces recherches, nombreuses et largement financées par des subsides publics aux Etats-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 évalués. 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 campagnes d'évaluation visent alors à remédier à ces différents problèmes en développant des tâches et des jeux de données publics, communs et réutilisables. Parallèlement des métriques automatiques sont mises au point afin de mesurer les performances, comparer les systèmes et leur évolution dans le temps. Les premières campagnes portent sur la compréhension de texte (conférences MUC, *Message Understanding Conferences*, 1987-1998) et seront suivies d'autres sur des thèmes similaires (recherche d'information, résumé automatique, etc.).

Les premières campagnes, qui avaient un rôle exploratoire et laissaient une grande marge de manœuvre aux participants, montrent que la compréhension de textes est 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 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.

2.1.2 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 repris 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 extensions, 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 texte : on peut ainsi

affecter à une même entité l'ensemble des informations qui la concerne, quelle que soit la forme sous laquelle cette entité a été dénommée en pratique.

- **Analyse des relations entre entités** : cette tâche, au nom explicite, vise à identifier les relations entre entités telle qu'elles sont exprimées dans les textes. L'analyse de relation 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 et pouvant être assimilés à des événements. Un événement dans ce cadre est donc défini comme un ensemble de relations entre entités identifiées au sein d'un texte donné.

D'autres modules peuvent bien entendus 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 impliquant des développements 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.

2.1.3 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 connaissance et de représentation très fouillés et, du coup, très coûteux à mettre en œuvre et peu portables d'un domaine à l'autre.
- les années 1990 voient quant à elle 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 montré que cette technologie n'était pas suffisante pour représenter toute la complexité des langues humaines mais plusieurs équipes montrent au cours des années 1990 que cette technologie est 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 à base d'apprentissage. Ces systèmes sont encore 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. 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 techniques 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 notamment ;

les systèmes à base d'apprentissage artificiel sont quant à eux beaucoup plus difficiles à corriger et faire évoluer[3]).

2.2 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'attarde 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 large 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.

3 Résumé automatique de textes d'opinion

Afin d'illustrer les enjeux de l'analyse automatique de textes en langage naturel, nous revenons ici sur une expérience passée concernant la production de résumés automatiques reflétant les opinions exprimées dans les textes. Cette expérience est intéressante car elle mêle analyse objective et subjective et montre bien, à notre avis, les limites des systèmes actuels.

Le point de vue sera épistémologique plus que technique : nous ne donnons pas, volontairement, tous les détails techniques qui ont déjà été discutés ailleurs (voir par exemple Bossard *et al.*, 2010 ou Poibeau, 2011). Notre but ici consiste davantage à examiner les possibilités techniques offertes, leur adéquation aux besoins et leurs limites prévisibles.

3.1 Problématique

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 les

[3] Notons qu'il s'agit d'un problème difficile car la complexité des systèmes à base de règles les rend eux-aussi difficile à faire évoluer manuellement. Globalement, la maintenabilité des systèmes de TAL reste un sujet de recherche aujourd'hui, surtout en milieu industriel.

mettre en avant et de jouer le rôle de « filtre » pour accéder rapidement à l'information pertinente.

Les techniques de résumé automatique ont beaucoup évolué ces dernières années afin de tenir compte des grandes quantités de texte disponibles. Deux évolutions majeures nous intéresseront ici : le résumé porte le plus souvent sur une grande masse de documents dont il faut faire la synthèse (par opposition au résumé mono-document, qui ne s'intéresse qu'à un document à la fois). Le résumé multi-documents laisse donc apparaître des problématiques particulières, au premier rang desquelles la reconnaissance des faits importants exprimés (par opposition aux faits secondaires ; autrement dit, il faut hiérarchiser l'information selon son importance) et la détection de la redondance (reconnaissance du fait que deux documents différents rapportent le ou les mêmes faits, éventuellement avec des information complémentaires ou contradictoires).

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 dans le cadre d'un système de questions réponses.

Le fonds documentaire était constitué de blogs et l'enjeu était de produire des synthèses cohérentes à partir de questions en langage naturel – en général, un résumé correspond à plusieurs questions liées (appelées *squishy list*) sur un thème donné (appelé *target*). Pour prendre un exemple, un des thèmes proposés visait la personnalité de l'année désignée par le magazine *Time* pour 2005 ("*Time Magazine 2005 Person of the Year*"). Les questions liées étaient les suivantes : "*Why did readers support Time's inclusion of Bono for Person of the Year?*", "*Why did readers not support the inclusion of Bill Gates as Person of the Year?*", "*Why did readers not support the inclusion of Melinda Gates as Person of the Year?*". On voit qu'il s'agit de questions en « pourquoi » (*why*) : contrairement aux questions factuelles (questions dites factoïdes, où la réponse est généralement une entité nommée), il n'est pas possible de répondre de façon simple à ces enchaînements de questions en pourquoi. Les systèmes de questions-réponses traditionnels, qui produisent des fragments en guise de réponse (*snippets*) sont insuffisants dans ce cadre, dans la mesure où ils ne permettent pas de « contextualiser » correctement la réponse, c'est-à-dire de produire un tout cohérent rendant compte des opinions exprimées. La production de résumés à partir d'une extraction de phrases donnant une idée des informations essentielles contenues dans le fonds documentaire et formant autant que possible un tout cohérent, semble une voie plus prometteuse.

3.2 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 retenue pour le résumé consiste à enrichir les documents avec différents types d'annotation, afin de normaliser les phrases et identifier celles qui rapportent le même type d'information. 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).

Une chaîne de traitement a alors été mise en place pour permettre une représentation de haut niveau du contenu des phrases. Pratiquement, la chaîne de traitement comportait les éléments suivants :

- découpage des textes en phrases et en mots ;
- annotation morphosyntaxique (identification de la catégorie de chaque mot et de ses principales caractéristiques : par ex. nom féminin singulier, verbe au passé simple de l'indicatif, etc.) ;
- annotation et catégorisation de certains termes clés (notamment pour les entités nommées : par ex. repérage que *France* est un nom de pays, que *Jacques Chirac* est un nom de personne, etc.) ;
- normalisation des termes repérés et résolution de la coréférence (par ex. repérage du fait qu'un pronom « *il* » réfère en fait à *Jacques Chirac*, que les séquences « *Chirac* » et « *le Président* » réfère à la même personne, etc.) ;
- repérage des verbes et autres éléments relateurs (par ex. *dissoudre* ou le nom prédicatif correspondant, *dissolution*) afin de permettre la mise en relation des différents éléments du texte.

Idéalement, ce type d'analyse devrait permettre de repérer directement des événements proches ou identiques. En fait, atteindre ce dernier niveau (c'est-à-dire le repérage de représentation formelles et normalisées de la notion d'événements) reste trop complexe pour l'état de l'art actuel, ce qui se manifeste par un silence important si l'analyse se limite au repérage de structures complexes formelles de haut niveau (le silence exprime le taux d'éléments non repérés par le système). Les systèmes pallient ces défauts en procédant à un calcul de similarité souple entre phrases : le repérage d'événements similaires ne repose pas tant sur le repérage de structures formelles identiques que sur un calcul statistiques prenant en compte le nombre d'éléments identiques ou similaires entre phrases (et s'appuyant donc essentiellement sur la catégorisation des éléments de la phrases en grandes classes sémantiques). Différentes formules statistiques ont été proposées pour le calcul de la similarité entre phrases (il s'agit d'un secteur de recherche florissant notamment dans le cadre de la recherche d'information) mais qui dépasse le propos de cet article. Nous laisserons le lecteur se référer à d'autres publications centrées sur ce problème (Achananuparp *et al.*, 2008).

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 juste 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.

Pour cela, nous avons assemblé des textes représentatifs (par exemple des textes issus de blogs ou les auteurs expriment directement leur état d'esprit positif ou négatif par le choix d'une icône particulière) pour repérer le vocabulaire essentiel pour l'expression d'opinions dans les domaines visés. Il est alors possible d'entraîner un « classifieur », c'est-à-dire de produire un programme capable de classer les textes voire les phrases selon l'opinion exprimée (pour les détails de l'algorithme et de la stratégie d'apprentissage, voir, Bossard *et al.*, 2010).

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'agir de résumé multi-documents, plusieurs phrases de différents documents source 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 peuvent être 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).

On obtient ainsi des textes représentatifs du corpus d'analyse, faisant un résumé des données orienté par la question posée a priori (cf. *supra* les notions de *target* ou de *squishy list*). Ces résumés peuvent ensuite être évalués suivant différents critères.

3.3 Résultats

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.

L'évaluation automatique reste donc un sujet de recherche en soi mais plusieurs métriques ont déjà été proposées. Il s'agit généralement de disposer d'un (ou souvent plusieurs) résumés de référence élaborés manuellement puis de comparer les résumés produits automatiquement avec ce(s) résumé(s) modèle. Le processus de comparaison des deux textes à des fins d'évaluation peut sembler très sommaire : il s'agit généralement de mesurer le nombre de séquences communes entre les textes (on voit qu'on est ici dans un processus purement mécanique, n'incluant aucune sémantique et encore moins une connaissance globale du contenu du texte). Il est aussi intéressant de constater que les mécanismes mis en œuvre vont au rebours de tous les présupposés nécessaires à l'évaluation d'une tâche aussi complexe que le résumé automatique, à savoir qu'il faut d'abor prendre connaissance globalement du texte, d'analyser les phrases et d'en donner une représentation sémantique. L'évaluation automatique évacue toutes ces problématiques en ayant simplement recours à la recherche des

séquences locales composées de *n* mots, sur une base purement mécanique et sans aucune connaissance du contenu.

L'évaluation automatique présente pourtant plusieurs avantages et conserve un grand intérêt pour tous les chercheurs du domaine :

1. L'analyse est automatique, elle ne coûte donc rien et est reproductible : la même métrique avec les mêmes donnés donnera toujours le même résultat, alors que les jugements humains varient de façon non négligeable d'une fois sur l'autre.
2. Elle est bien corrélée aux évaluations manuelles. C'est-à-dire qu'une note en soi ne veut pas dire grand chose et doit être interprétée avec précaution. En revanche un ensemble de notes peu permettre de comparer différents analyseurs (ou différentes versions d'un même analyseur) et les jugements relatifs (du système A vis-à-vis du système B ou de telle version du système A par rapport à telle autre version du même système) sont très souvent conformes aux résultats obtenus avec des évaluations manuelles.

Lors de la campagne 2008, les deux types d'évaluation ont été mis en œuvre. D'une part une analyse automatique donnait un indice de la qualité relative des différents systèmes de résumé. D'autre part plusieurs indices difficiles à évaluer automatiquement faisaient l'objet d'une évaluation par des experts humains (grammaticalité, lisibilité, cohérence, non redondance, ainsi qu'un dernier indice estimant dans quelle mesure le texte produit dans le résumé répondait à la question posée au départ).

Le système que nous avions 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 proposé, 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.

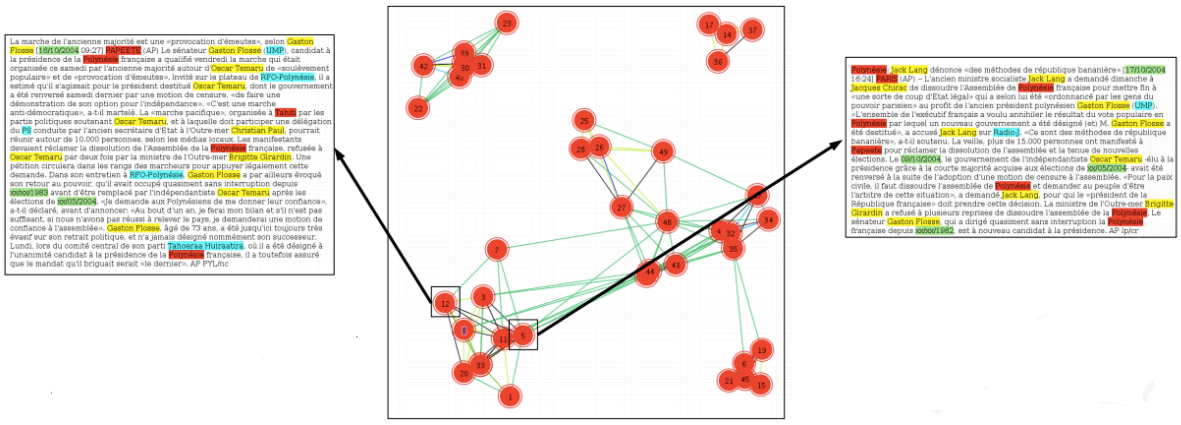

Figure 1 : regroupements de documents en classes événementielles. Chaque point renvoie à un document qui peut être visualisé à la demande. La même représentation est possible au niveau des phrases.

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.

3.4 Discussion

Cette expérience nous semble intéressante à plusieurs titres : elle montre qu'il est désormais possible d'analyser automatiquement de grandes masses de données de manière relativement fine. Cette analyse automatique ne se substitue pas à une analyse humaine mais peut l'assister efficacement pour traiter notamment la quantité de documents quand celle-ci dépasse les capacités de lecture du chercheur ou de l'analyste.

Il faut malgré tout noter les limites des systèmes automatiques. Même quand l'information est locale et clairement exprimée, l'ordinateur fait des erreurs fréquentes et parfois grossières. Ce point pose des problèmes non résolus à ce jour : les ordinateurs peuvent traiter de grandes masses de données mais quid des erreurs ? des oublis ? des signaux faibles ? Autant de points fondamentaux qui, suivant les applications considérés, peuvent poser problème si on s'en tient à une analyse automatique.

Les résultats restent aussi mitigés dès que l'analyse suppose une évaluation de l'importance et de la pertinence de l'information recherchée. Cette évaluation suppose 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.

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 fine et correcte (en revanche, une analyse en terme 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 laissés de côté jusqu'à

récemment, la priorité ayant longtemps été donnée aux informations dites factuelles. L'avènement du Web participatif, notamment des forum et des avis de consommateurs, a eu un poids considérable pour le renforcement des recherches en ce domaine.

Les éléments linguistiques pris en compte sont aussi intéressants. Il s'agit essentiellement des adjectifs, des adverbes voire des négations, des éléments largement passés de côté jusque là (la négation avait suscité de très nombreuses recherches en linguistique formelle mais ces recherches son très éloignées des besoins pratiques et des techniques développées actuellement face à des problèmes très concrets. Les ressources développées pour 'l'analyse d'opinion sont par ailleurs sans doute utiles, mais elles sont souvent peu adaptées ou très incomplètes : suivant la source d'information, l'auteur et l'objet considéré (par ex. s'il s'agit de revues de produits sur le Web), le vocabulaire employé sera très différent. Les techniques d'apprentissage ont là encore donné des résultats remarquables en permettant l'analyse dynamique de corpus particuliers afin d'identifier le lexique le plus spécifique (positif et négatif) et donc le plus pertinent pour un domaine ou une source donnée. On a ainsi des systèmes facilement adaptables et évolutifs en fonction des données.

4 Enjeux pour les sciences sociales : l'exemple de la campagne PoliInformatics 2014

On assiste depuis quelques 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 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?*" or "*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'informations sur les éléments saillants de la crise financières (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 entités et de naviguer au sein du réseau ainsi obtenu. Une autre représentation utile des données consiste à voir l'évolution des thèmes au cours du temps.

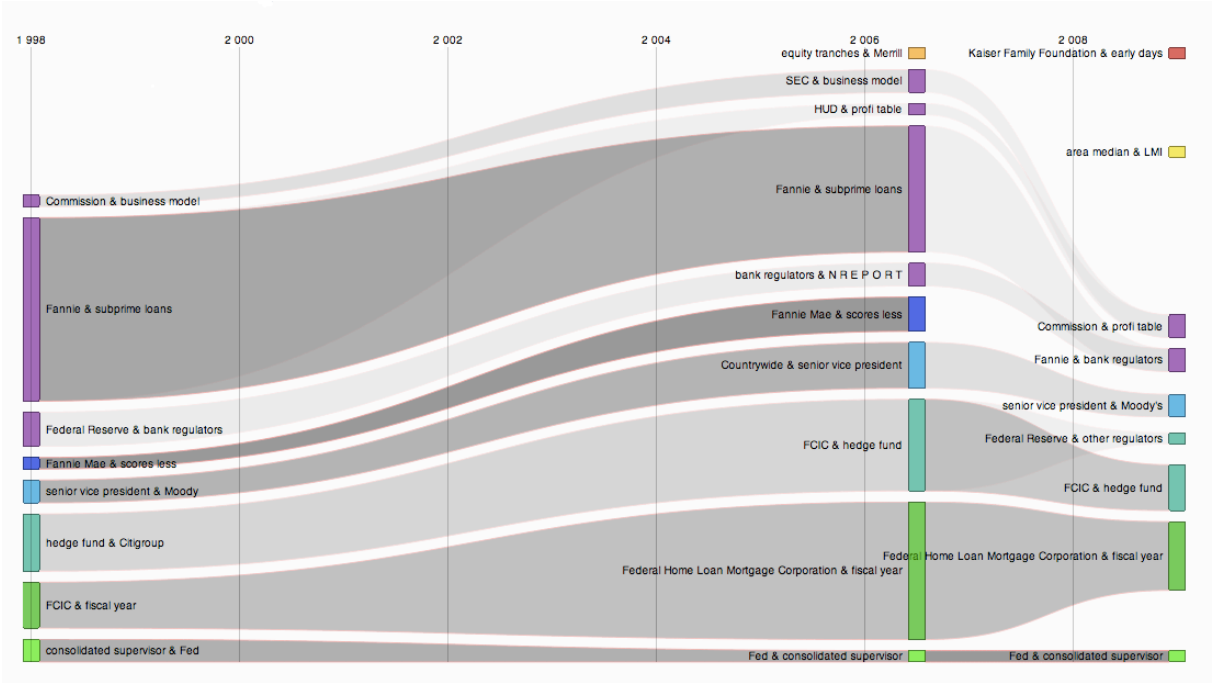

Figure 2 : identification de thèmes et visualisation de leur évolution au cours du temps dans une partie du corpus PoliInformatics. Cette visualisation est obtenue grâce à la plate-forme Cortext de l'IFRIS.

A 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 acteur par rapport aux faits évoqués. Les outils automatiques sont donc utiles mais ils ne peuvent faire qu'une partie 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.

L'interprétation des données reste en revanche entièrement à la charge de l'expert. On peut ici se rappeler ce que l'on a pu observer 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 d'autres ou vérifier la validité d'une source donnée. 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. Ils peuvent aussi fournir une vue globale sur un corpus par exemple, et différents moyens d'accès à travers des cartes et des outils de navigation interactifs.

5 Conclusions

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 diversifiés (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, Palakorn, Xiaohua Hu et Xiajiong Shen (2008). The Evaluation of Sentence Similarity Measures. *Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery* (Turin, Italy). Springer-Verlag.

AlTantawy Mohamed, Alix Rule, Owen Rambow, Zhongyu Wang et Rupayan Basu (2014). Using Simple NLP Tools to Trace the Globalization of the Art World. *Proc of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, USA.

Baccianella Stefano, Andrea Esuli et Fabrizio Sebastiani (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 Steven, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou et Dan Jurafsky (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 Aurélien et Thierry Poibeau (2008). Regroupement automatique de documents en classes événementielles. *Actes de la conférrence Traitement Automatique de Langage Naturel* (TALN 2008). Avignon.

Bossard Aurélien, Michel Généreux et Thierry Poibeau (2010). Résumé automatique de textes d'opinion. *Traitement Automatique des Langues*, 51/3, pp. 47-73.

Bourreau Pierre et Thierry Poibeau (2014). Mapping the Economic Crisis: Some Preliminary Investigations. *Technical report on the ACL 2014 PoliInformatics Unshared task*. arXiv:1406.4211.

Dey Lipika et Mirajul Haque (2008). Opinion mining from noisy text data. *AND 08: Proceedings of the second workshop on Analytics for noisy unstructured text data*. NewYork.

Esuli Andrea et Fabrizio Sebastiani (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of the conference on Language Resources and Evaluation* (LREC'06). Gènes.

Gaussier, Eric et François Yvon (2011). *Modèles statistiques pour l'accès à l'information textuelle*. Lavoisier (Coll. Recherche d'information et web), Paris.

Hobbs, Jerry R., Douglas E. Appelt, John Bear, David Israel, Megumi Kameyama, et Mabry Tyson (1993). FASTUS: A System for Extracting Information from Text. *Proceedings of the Human Language Technology Conf.*, Princeton, New Jersey, pp. 133-137.

Hutchins, John (2001). Machine translation over fifty years. *Histoire, Epistémologie, Langage.* Vol. 23 (1), 7-31.

Kao Huan-An, Hsin-Hsi Chen (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, Vasileios, Daniel Preotiuc-Pietro, Sina Samangooei, Douwe Gelling et Trevor Cohn (2014). Extracting Socioeconomic Patterns from the News: Modelling Text and Outlet Importance Jointly. *Proc of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, USA.

Poibeau, Thierry (2003). *Du texte brut au web sémantique*. Lavoisier, Paris.

Poibeau, Thierry (2011). *Traitement automatique du contenu textuel*. Lavoisier, Paris.

Pak Alexander et Patrick Paroubek (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.

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

Strapparava Carlo et Alessandro Valitutti (2004). WordNet-Affect : an affective extension of WordNet. *Proceedings of the conference on Language Resources and Evaluation* (LREC'14). Lisbonne., p. 1083-1086.

Tellier, Isabelle et Mark Steedman (2010). Apprentissage automatique pour le TAL. *Traitement Automatique des Langues* (TAL), 50/3.

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

Venturini, Tommaso et Daniele Guido (2012). Once Upon a Text: an ANT Tale in Text Analysis. *Sociologica* (Italian Journal of Sociology Online), 2012/3.

Vernier Matthieu et Laura Monceaux (2007). Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques. *Traitement Automatique des Langues*, 2007.

Vlachos, Andreas et Sebastian Riedel (2014). Fact Checking: Task definition and dataset construction. *Proc of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, USA.

Wiebe Janyce, Theresa Wilson et Matthew Bell (2001). Identifying Collocations for Recognizing Opinions. *Proc. Of the ACL 2001 Workshop on Collocation*. Toulouse.

Zhang Wei, Shuang Yu et Weiyi Meng (2007). Opinion retrieval from blogs. *CIKM '07 : Proceedings of the sixteenth ACM Conference on Information and Knowledge Management*. NewYork.