Compte rendu de l'atelier Journalisme computationnel 2017

Laurent Amsaleg, Vincent Claveau

Compte rendu de l'atelier Journalisme computationnel 2017

Notes about the Computational Journalism workshop 2017

Laurent Amsaleg, Vincent Claveau

CNRS, IRISA, Campus de Beaulieu, Rennes, France
prenom.nom@irisa.fr

RÉSUMÉ. Ce billet dresse un bilan des présentations et des discussions qui ont eu lieu lors de l'atelier « Journalisme computationnel » du 24 janvier 2017. L'atelier était organisé par Laurent Amsaleg (CNRS, IRISA), Vincent Claveau (CNRS, IRISA) et Xavier Tannier (LIMSI-Univ. Paris Sud). Il était adossé la conférence EGC2017 se tenant à Grenoble.

ABSTRACT. This short paper gives an overview of the presentations and discussions held during the "Computational Journalism" workshop. This workshop was proposed by Laurent Amsaleg (CNRS, IRISA), Vincent Claveau (CNRS, IRISA) and Xavier Tannier (LIMSI-Univ. Paris Sud). It took place during the EGC2017 conference in Grenoble, France.

MOTS-CLÉS : Journalisme computationnel, journalisme de données, vérification des faits, traitement automatique des langues.

KEYWORDS: Computational journalism, data journalism, fact checking, natural language processing.

1 Motivations et objectifs de l'atelier

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

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

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

1. http://compjournalism2016.irisa.fr

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

Autour de cette volonté de croiser les discours, de faire des allers-retours entre journalistes et informaticiens, nous avons bâti un appel à communications dont le cœur est constitué des thèmes suivants : détection d'événements, vérification *(fact-checking)*, études sociologiques ou historiques, exploration d'archives d'actualités, génération automatique de contenu journalistique, visualisation de données, navigation dans de grandes masses de données, myriadisation *(crowdsourcing)* pour le journalisme, dissémination des nouvelles à travers les réseaux sociaux, contextualisation, recommandation, détection de plagiat, de cliché, de biais, de propagande, de fausses informations *(hoax, fake news)* dans le texte, les images ou les vidéos...

2 Programme

Les articles reçus en réponse à l'appel ont été relus par un comité de lecture incluant informaticiens et professionnels de l'information. Ils nous ont ainsi permis de bâtir un programme varié autour de trois exposés longs et cinq exposés courts abordant des points techniques, livrant des analyses sociologiques, ou témoignant des usages journalistiques d'outils informatiques (Amsaleg *et al.*, 2017).
Exposés longs :

- Fouille de textes pour une analyse comparée de l'information diffusée par les médias en ligne : une étude sur trois éditions du Huffington Post (Velcin *et al.*, 2017)
- Génération automatique de billets journalistiques : singularité et normalité d'une sélection (Vizzini *et al.*, 2017)
- Analyse des média français : quand l'économie rencontre la fouille de donnée (Viaud *et al.*, 2017)

Exposés courts :

- Étude des influences réciproques entre médias sociaux et médias traditionnels (Mazoyer *et al.*, 2017)
- Analyse exploratoire de corpus textuels pour le journalisme d'investigation (Médoc *et al.*, 2017)
- Détection automatique de grandes thématiques de la propagande Nord Coréenne (Grabar & Richey, 2017)
- Utilisation conjointe LDA et Word2Vec dans un contexte d'investigation numérique (Maitre *et al.*, 2017)
- Erreurs OCR et biais d'indexation : impact sur les usages (Chiron *et al.*, 2017)

Pour compléter cette sélection, nous avons invité Pierre Bellon, Gauthier Bravais et Lucas Piessat, de l'agence Skoli, pour présenter leurs analyses du traitement médiatique de faits d'actualités mêlant utilisation d'outils de fouille et rendu grand-public. Leur présentation était intitulée : « Une analyse de données textuelles des archives numériques de la presse française pour explorer le traitement médiatique de l'Islam. L'exemple d'une collaboration chercheur / agence Web spécialisée ». Dans cette présentation, ils nous ont expliqué leur travail conjoint avec Moussa Bourekba (chercheur CIDOB, Barcelone) pour étudier le traitement médiatique de l'Islam en France (1997-2015). Leur collaboration, originale à ce niveau, s'est articulée autour d'une analyse de données textuelles de milliers d'articles issus des archives numériques des journaux Le Monde, Le Figaro et Libération, et de sa restitution par une interface Web mêlant data-visualisations et décryptages.

3 Un premier constat : le paysage change

Les participants, une vingtaine, font tous le constat que le paysage médiatique moderne est en plein bouleversement, et plusieurs discussions mettent en avant les nouvelles pratiques que l'on observe. Il est constaté que les manières de produire de l'information, au sens assez large, se démarquent des pratiques plus anciennes. L'information n'est plus uniquement issues des producteurs historiques (télévision, organes de presse), mais vient aussi des citoyens, ce qui pose, d'ailleurs, des problèmes de véracité et de confiance. Non seulement la nature des sources d'information évolue, maintenant éparses, multiples, mais cette information n'est plus uniquement relayée par les producteurs historiques. Web et réseaux sociaux se posent comme autant de relais, parfois plus réactifs.

L'audience change également : elle se morcelle, et n'attend plus que l'information soit délivrée à des moments établis (journal télévisé, éditions quotidiennes pour les journaux). La consommation d'information est désormais en flux ; chacun peut piocher à tout moment. Cette consommation est très ubiquitaire, s'adapte au rythme de chacun *(replay, podcasts)*, avec des lectures différenciées selon les supports, les moments de la journée, les contextes (transport en commun, lecture au calme. . .). Chacun espère une certaine forme de personnalisation, en mettant en avant quelques sujets au détriment d'autres ou encore en permettant de choisir une forme, une présentation qui soit perçue comme la plus agréable possible. En plus de ces canaux informationnels, se créent en parallèle d'autres flux consacrés par exemple au *fact checking*, à l'enrichissement de contenus sur second écrans, etc. Il est également observé que ces flux, ou encore ces analyses du matériau informationnel sont destinés au grand public et que, par conséquent, l'information doit être claire, intelligible, et possiblement interactive, et qu'une certaine sobriété est de mise pour espérer un abord facile. Tous les participants à l'atelier constatent la profonde transformation numérique du monde qui entraîne notamment une surabondance d'informations. Et tous espèrent que les éléments discutés ici participeront à en faciliter l'accès, à mieux en établir la véracité ou au contraire à pointer du doigt leur caractère mensonger.

4 Le journalisme computationnel pour qui ?

Deux grandes familles de besoins d'analyse de l'information se sont dégagées des discussions durant l'atelier. Tout d'abord, les journalistes eux-mêmes ont besoin d'outils d'analyse. D'une part, les outils actuels sont parfaitement adaptés pour soulager les rédactions de la création d'articles répétitifs, très factuels (résultats d'élections ou de rencontres sportives). Il n'y a pas d'interprétation de l'information ici, juste une mise en forme plaquée sur un canevas, mise en forme qui peut être assez sophistiquée et offrir une certaine diversité dans le texte créé. D'autre part, les journalistes ont besoin d'outils pour faciliter leur compréhension de larges volumes d'informations. Certaines techniques, par exemple, sont aptes à faire émerger d'une large collection de données textuelles des thématiques méritant peut être examen plus approfondi. Il y a là une volonté de cartographier les contenus, de les regrouper ou de les séparer, afin de fournir autant d'hypothèses que les journalistes visualiseront et vérifieront ensuite. Ces outils s'appuient par exemple sur la génération de petits nuages de mots, fournissent les différentes variantes des sujets découverts, permettent de visualiser les documents associés selon plusieurs niveaux de détails. Ces outils semblent être intéressants dans le contexte du journalisme d'investigation, par exemple. Une autre étude vise par exemple à faire émerger les thématiques de la propagande Nord Coréenne (Grabar & Richey, 2017), et permet d'établir que le registre de langage très belliqueux vis-à-vis de l'extérieur est perçu positivement de l'intérieur.

Par ailleurs, les sociologues ont aussi besoin d'outils d'analyse mais plutôt pour comprendre ce qu'est l'information, comment elle évolue, etc. Par exemple, Vizzini *et al.* (2017) proposent des outils pour comparer les éditions internationales du même journal afin de faire ressortir les particularités de chaque pays qui se matérialisent par les sujets traités ou ignorés, et par l'importance relative de chaque sujet au sein de chaque édition. Certaines catégories d'informations n'existent que dans un pays. Cette même étude compare les catégories regroupant les informations créées d'une part par les journalistes et les catégories qui émergent de l'analyse automatique du corpus des articles au travers de regroupements thématiques. On voit que les catégories diffèrent, celles des rédactions ayant parfois une dimension d'accroche, en décalage avec le contenu informationnel.

D'autres études visent à caractériser les manières selon lesquelles les informations circulent entre les médias (Viaud *et al.*, 2017). Qui a la primauté de l'information, qui la reprend, en quels termes, comment cela traverse-t-il

les réseaux sociaux ? Une des questions à la racine de cette étude était de savoir si un média tire vraiment avantage à lancer une information originale. Il en ressort qu'il existe une forte corrélation entre la taille d'une rédaction et sa capacité à produire de l'information en premier, et que plus que d'être étiqueté comme lanceur, c'est le sérieux de média qui s'établit graduellement.

5 Quels matériaux, quelles approches ?

La majeure partie des outils présentés lors de l'atelier se fonde sur une analyse du texte. Comptage de mots, regroupement en thèmes, extraction d'entités nommées, mise en relation avec des ontologies, on retrouve ici le corpus technologique traditionnel de la fouille de texte. De ce point de vue technique, on note en particulier l'utilisation dans beaucoup des travaux présentés (Velcin *et al.*, 2017; Maitre *et al.*, 2017; Grabar & Richey, 2017, *inter alia*) de méthodes de découverte avec des approches non-supervisées : *clustering*, LDA, plongements de mots *(word2vec)*... Il convient aussi de noter que les textes utilisés peuvent avoir des origines et des caractéristiques différentes : agences et journaux, traditionnels ou *pure-player* (Viaud *et al.*, 2017), de plusieurs pays (Velcin *et al.*, 2017), textes collectés sur le Web (Grabar & Richey, 2017), sur les réseaux sociaux (Mazoyer *et al.*, 2017)...

L'absence de traitements d'autres modalités, comme l'image, la vidéo, est cependant à noter. L'analyse d'images, la mise en relation d'images identiques présentes dans de multiples textes, la détection d'images falsifiées ont un intérêt évident dans la thématique du journalisme computationnel. Ce champ est actif, comme en témoignent les projets InVid (`www.invid-project.eu/consortium/`) ou Maven (`http://maven-project.eu/`). L'absence de travaux s'y rapportant lors de l'atelier est donc regrettable, tout comme l'absence des aspects représentation des connaissances, Web sémantique et visualisation.

6 Un domaine plein de défis

6.1. Difficultés scientifiques

Les échanges entre les participants font ressortir plusieurs difficultés conceptuelles ou techniques qui brident à l'heure actuelle nos capacités à proposer des outils participant au journalisme computationnel. Tout d'abord, le problème de la représentation de l'information reste entier. L'information est diverse, constituée d'archives plus ou moins bien structurées, stockées selon des formats hétérogènes. Ces informations ont souvent besoin d'être croisées avec des sources externes, issues par exemple d'institutions compilant des statistiques ouvertes sur tel ou tel sujet. Or, nos algorithmes ont pour l'instant besoin de travailler avec des données représentées plutôt de manière uniforme pour en faciliter les traitements. RDF *(Resource Description Framework)*, graphes ou autres représentations sont autant d'options possibles, avec chacune des bons et des mauvais côtés. Il n'y a pas de consensus, chacun se débrouille pour créer des passerelles entre ces multiples formats, ce qui reste peu satisfaisant. De plus, les besoins de représenter l'information, tout du moins logiquement, selon différents niveaux de détails sont importants, et on ne sait pas clairement comment faire.

À côté de ces écueils liés à la représentation existent aussi des difficultés liées à la prise en compte du temps qui s'écoule. Aucune des contributions présentées à cet atelier n'est capable de montrer comment évoluent au cours du temps les informations observées. Les travaux prennent des clichés figeant un instant ou une période, mais sont incapables de mettre en lumière les évolutions dans le temps. De plus, tous les participants s'accordent à dire que cette évolution est complexe car multiforme, on veut pouvoir parler des évolutions temporelles des informations selon différents angles de vue. Personne ne sait très clairement comment représenter la temporalité bien que le besoin de le faire soit partagé.

6.2. Difficultés d'usage

Un constat unanime des participants est que nous avons à faire face à des difficultés liées à l'acceptation par les journalistes, les sociologues, etc., de ces technologies. Ces difficultés ont plusieurs causes. Tout d'abord, la compréhension des résultats est un point important pour l'acceptabilité d'une technologie. Il se peut qu'il nous

faille accompagner les résultats produits par nos algorithmes de certaines informations expliquant pourquoi telle ou telle décision a été prise, pourquoi ces informations-ci sont regroupées avec celles là. Bien entendu, il ne s'agit pas de noyer les utilisateurs sous des explications techniques, mais de trouver les éléments qui leur donnent confiance dans les résultats. Ceci est probablement très difficile à faire.

De plus, les délais entre le besoin journalistique et sa concrétisation technologique sont peu compatibles. Il est naturel que les utilisateurs (journalistes ou sociologues) aient envie de tester tout de suite ceci ou cela, ce qui est souvent incompatible avec le temps nécessaire pour transformer une idée en système opérationnel. Ce décalage frustrant est exacerbé par la culture de l'instantané, de l'actualité chaude. Les échanges des chercheurs avec les journalistes indiquent que ces derniers préfèrent utiliser des outils naïfs, mais qui sont immédiatement disponibles et dont ils savent se servir, pour traiter un événement d'actualité plutôt que de travailler plus en profondeur, ce qui les ferait s'éloigner du temps fort. Plus largement enfin, les professionnels de l'information, de part leurs itinéraires de formation, ont rarement les compétences adéquates pour comprendre et utiliser les technologies que les chercheurs peuvent proposer.

7 Bilan

Cet atelier a ainsi permis de dresser un panorama assez diversifié, même si nous le savons incomplet, des usages journalistiques envisagés et des techniques utilisées. Il a aussi permis de recenser un certains nombres de défis, scientifiques ou non, à aborder pour développer des technologies réellement pertinentes pour le journalisme computationnel. Signalons enfin que d'autres appels sont en cours pour des sessions spéciales aux conférences WIFS et Multimedia Modeling 2018. Elles visent plus spécifiquement certaines communautés citées en Section 5 et se placent dans un cadre cette fois-ci international.

Références

L. Amsaleg, V. Claveau & X. Tannier, Eds. (2017). *Actes de l'atelier Journalisme computationnel, conjoint à la conférence EGC*, Grenoble, France.

Chiron G., Moreux J.-P., Doucet A., Coustaty M. & Visani M. (2017). Erreurs OCR et biais d'indexation : impact sur les usages. In (Amsaleg *et al.*, 2017).

Grabar N. & Richey M. (2017). Détection automatique de grandes thématiques de la propagande nord coréenne. In (Amsaleg *et al.*, 2017).

Maitre J., Menard M., Chiron G. & Bouju A. (2017). Utilisation conjointe LDA et Word2Vec dans un contexte d'investigation numérique. In (Amsaleg *et al.*, 2017).

Mazoyer B., Turenne N. & Viaud M.-L. (2017). Étude des influences réciproques entre médias sociaux et médias traditionnels. In (Amsaleg *et al.*, 2017).

Médoc N., Ghoniem M. & Nadif M. (2017). Analyse exploratoire de corpus textuels pour le journalisme d'investigation. In (Amsaleg *et al.*, 2017).

Velcin J., Soulages J.-C., Kurpiel S., Otavio L., Vecchio M. D. & Aubrun F. (2017). Fouille de textes pour une analyse comparée de l'information diffusée par les médias en ligne : une étude sur trois éditions du huffington post. In (Amsaleg *et al.*, 2017).

Viaud M.-L., Hervé N. & Cagé J. (2017). Analyse des media français : quand l'économie rencontre la fouille de donnée. In (Amsaleg *et al.*, 2017).

Vizzini J., Labbé C. & Portet F. (2017). Génération automatique de billets journalistiques : singularité et normalité d'une sélection. In (Amsaleg *et al.*, 2017).