LE TOUR DE LA QUESTION

« Robot-journalisme » et production automatisée de contenus : bilan des premières initiatives et perspectives pour les médias

Après les algorithmes conçus pour prédire les sujets qui vont faire l'actualité en ligne, des algorithmes dédiés à la production automatisée de contenus journalistiques apparaissent. Plusieurs outils, proposant de transformer des données brutes en langage « humain » sont déjà utilisés par des rédactions, essentiellement américaines, pour des contenus relatifs au sport, à l'économie et aux finances. En parallèle, de nouveaux acteurs utilisent également ces technologies pour développer de nouveaux produits dans une approche BtoC. Quelles sont les start-up positionnées sur ce secteur et comment fonctionnent leurs outils ? Comment sont-ils utilisés par les rédactions ? Quels sont les résultats obtenus et ceux envisagés ? Après avoir présenté les différents outils et projets, Satellinet dresse un premier bilan des initiatives mises en place et évoque les perspectives offertes pour les médias.

Le « robot-journalisme », est le terme utilisé pour désigner **la production automatisée ou semi automatisée de contenus journalistiques, grâce des algorithmes permettant de transformer directement des données brutes en texte**. Souvent abordées sous l'angle des répercussions potentielles pour l'emploi dans le domaine du journalisme, ces solutions sont au contraire présentées par leurs créateurs comme des outils censés donner plus de valeur au travail journalistique. Pour les éditeurs, ils représentent une opportunité d'augmenter la production éditoriale tout en limitant ses coûts.

« Le but est de sous-traiter les tâches les plus rébarbatives à des machines afin d'enrichir le travail des rédactions. **Il ne faut pas aborder ces outils comme des éléments qui vont remplacer le journaliste**. Le journaliste doit donner du sens, par exemple trouver ce qui n'est pas divulgué dans un rapport financier, ce qu'un algorithme n'est pas capable de faire », estime Eric Scherer, directeur de la prospective directeur de la prospective, de la stratégie numérique et des relations internationales liées aux nouveaux médias à France Télévisions.

« L'objectif est de sous-traiter les tâches les plus rébarbatives à des machines afin d'enrichir le travail des rédactions »

Eric Scherer (France Télévisions)

Lancés depuis 2012 aux Etats-Unis, les premiers projets sont opérationnels depuis quelques mois. Ils reposent sur des partenariats avec des sociétés spécialisées dans la R&D et dans le traitement automatique du langage (TAL), discipline au croisement de la linguistique, de l'informatique et de l'intelligence artificielle. Ces dernières fournissent aux éditeurs des plateformes technologiques permettant de générer des textes de façon automatique, dans une démarche BtoB.

LES EXPÉRIENCES AMÉRICAINES DANS LA GÉNÉRATION D'ARTICLES

Fondée en 2007 par Robbie Allen, la société américaine Automated Insights, fait souvent figure de pionnier dans la production de contenus automatisée. Basée à Durham (Etats-Unis), la société a levé, depuis sa création, 10,8 millions de dollars (8,7 millions d'euros) auprès de divers fonds d'investissement (IDEA Fund Partners, OCA Ventures, Court Square Ventures, Samsung Ventures, Osage Ventures Partners, Valahlla Partners, Box Group), d'investisseurs privés, mais aussi de l'agence de presse américaine Associated Press (AP), en 2014. La société compte 35 collaborateurs, répartis entre commerciaux, experts en technologies, et linguistes. Automated Insights compte parmi ses principaux clients Yahoo!, Associated Press (AP), La NFL (la fédération nationale de football américain), Samsung, et Edmunds.com (site dédié à l'automobile).

« Pendant de nombreuses années, nous avons passé beaucoup de temps à extraire des chiffres et réécrire des informations provenant des rapports financiers. **Au lieu de fournir 300 articles manuellement, nous sommes désormais capables d'en produire 4 400 à chaque trimestre de façon automatique »,** explique Lou Ferrara, managing editor de AP. Cette automatisation repose sur la solution « Wordsmith » développée par Automated Insights,

>>>

LE TOUR DE LA QUESTION

>>>

permettant de générer des textes en langage « humain » à partir des données brutes fournies par Zacks Investment Research, fonds d'investissement et société spécialisée dans l'analyse financière, qui stocke toutes les données issues des documents financiers des entreprises américaines cotées.

Depuis juillet, des milliers d'articles de AP sont ainsi produits de façon automatique, et repris par les centaines d'éditeurs, abonnés au flux de l'agence. La plateforme permet une vérification manuelle avant publication et est paramétrée pour rédiger des articles dont le style est le même qu'une dépêche de l'AP. Plutôt que de remplacer les journalistes financiers, l'automatisation des contenus est censée leur permettre d'allouer davantage de temps et d'énergie à des articles à plus haute valeur ajoutée.

Le même argument est mis en avant par **Local Labs**, société basée à Chicago et positionnée sur la production d'informations hyperlocales en sous-traitance. **Elle a levé 4,6 millions de dollars (3,7 millions d'euros) auprès du groupe média Tribune Media Company** (Chicago Tribune, Los Angeles Times, Baltimore Sun, etc.). A l'occasion de cette prise de participation, The Chicago Tribune avait cependant supprimé environ la moitié des 40 postes de journalistes en charge des pages des informations locales pour la sous-traiter à cette société. Fondée en 2006 initialement sous le nom de Journatic Labs, la société a opéré en 2014 une opération de rebranding, en partie liée à plusieurs scandales de plagiat et d'utilisation de faux noms d'auteurs. Ces derniers avaient notamment poussé Mike Fourcher, directeur éditorial de la société, à démissionné en 2012.

La société revendique désormais 25 employés, auxquels s'ajoutent 100 collaborateurs indépendants, et produit pour ses clients des informations locales de base (informations pratiques, calendriers, brèves, annonces immobilières, résultats sportifs...), directement intégrables en formats papier ou web. La société commercialise aussi une offre de données brutes, baptisée « Community Data », et regroupant, par municipalité, différents types de données (annonces immobilières, résultats sportifs, permis de construire, rapports financiers, données publiques...).

Egalement basée à Chicago, **Narrative Science a mis au point la solution « Quill », plateforme permettant de transformer automatiquement des données ou des rapports financiers en courts textes intelligibles.** Fondée en 2010 comme un projet de la Northwestern University, Narrative Science a, depuis, levé 22,4 millions de dollars (18 millions d'euros), auprès notamment des fonds d''investissements Battery Ventures, Sapphire Ventures, Jump Capital, et InQ Tel. Parmi les principaux clients de cette solution figurent de grands groupes comme Mastercard, Deloitte et Publicis, à qui elle fournit des rapports simplifiés à partir de leurs données. **La société a également conclu des partenariats avec des éditeurs comme** Forbes, pour lequel elle produit automatiquement des articles à partir de rapports financiers, et le site ProPublica, pour lequel elle génere automatiquement de courtes descriptions des établissements scolaires dans le cadre d'un projet éditorial consacré aux inégalités d'accès à l'école.

« Quill écrit déjà pour les entreprises afin qu'elles puissent mieux comprendre leurs données. Bientôt, Quill pourrait être utilisé dans les salles de rédaction, pour faciliter la compréhension de vastes bases de données, et permettre aux journalistes des analyses plus approfondies », affirme Andrew Paley , responsable UX à Narrative Science.

Mis en place par Ken Schwencke, journaliste et développeur au Los Angeles Times, **Quakebot est quant à lui un algorithme relié directement à la base de données de l'US Geological Survey (Institut d'études géologiques des États-Unis).** Il est capable d'écrire presque en temps réel des articles relatifs aux tremblements de terre survenus en Californie. En développement depuis 2012, Quakebot est opérationnel depuis juillet 2014. « L'objectif est de pouvoir extraire les informations de base de façon la plus rapide et la plus précise possible, avant de décider lesquelles méritent un traitement plus approfondi », explique Ken Schwencke, interrogé par Slate. Le Los Angeles Times a par ailleurs mis en place le même genre de dispositif avec les rapports de police, publiant un bref compte-rendu et mettant à jour une carte en temps réel à chaque homicide .

LES INITIATIVES EN FRANCE

En France, plusieurs acteurs se positionnent sur le traitement automatique du langage, mais les initiatives concernent pour l'instant des applications autres que la génération d'articles. Les algorithmes sont notamment utilisés à des fins de retranscription, de valorisation de contenus et de recommandation éditoriale.

Depuis septembre 2014, **le SID (Sport-Informations-Dienst),** filiale allemande de l'AFP **expérimente la production automatisée d'articles sportifs,** via un partenariat avec la start-up aexea, qui propose une plateforme Saas de génération automatique de texte dans 13 langues, pour des contenus relatifs au sport, à la finance ou à la météo. Un projet pilote est en cours avec le SID pour transformer automatiquement des résultats sportifs bruts en brefs comptes rendus. Mais en France, aucun projet comparable n'a pour l'instant

>>>

LE TOUR DE LA QUESTION

>>>

été lancé. L'AFP dispose déjà d'un outil interne permettant de traiter des données et de générer automatiquement des résultats sportifs bruts, lesquels sont ensuite validés par les journalistes. « Nous n'avons pas lancé, à ce stade, de projets de génération d'articles sportifs (en français ou anglais), mais nous y réfléchissons avec des partenaires. **Nous travaillons actuellement sur plusieurs autres prototypes automatiques, notamment dans le domaine de la visualisation des informations et de la retranscription de discours à partir des pistes audio des vidéos** », indique Denis Teyssou, responsable éditorial du Medialab, pôle R&D de l'AFP.

Dans le domaine de la visualisation, le projet Earthnews de l'AFP permet par exemple l'affichage dynamique des informations multimédia géolocalisées de l'AFP sur une carte de la NASA, mise à jour en temps réel. Ce projet est conçu pour afficher de l'information en continu sur les écrans publics, notamment pour les endroits ne pouvant diffuser de son, comme les terminaux d'aéroports, les salles ou files d'attente...

La retranscription automatique de la parole (discours audio et vidéo) sous forme écrite permet quant à elle d'identifier plus rapidement les moments intéressants d'un discours, et d'améliorer le référencement par les moteurs de recherche des vidéos produites par l'agence. L'AFP a travaillé dans plusieurs projets de R&D sur la transcription de la parole en français, anglais et arabe, notamment avec la société Vocapia Research.

Fondée en 2000, Vocapia Research édite une suite logicielle (VoxSigma) dans le domaine du traitement automatique de la parole. La société de R&D, partenaire du laboratoire LIMSI (CNRS) compte une dizaine de collaborateurs, et prévoit de réaliser deux millions d'euros de chiffre d'affaires en 2014, dont 20 % dans le secteur des médias, et 30 % à l'international.

« Nos logiciels sont commercialisés sous forme de licences à installer sur une machine Linux, ou sous forme de service via une solution SaaS. Avec l'AFP, nous avons un partenariat en vue de développer un pilote de retranscription automatique de leurs vidéos. En échange, nous avons accès à leurs flux de dépêches afin de mettre à jour le vocabulaire de notre système, et en particulier de l'enrichir quotidiennement avec les nouveaux mots apparaissant dans l'actualité », explique Bernard Prouts, fondateur et CEO de Vocapia Research.

L'AFP est également en discussions avec **Syllabs**, start-up fondée en 2006 par Claude de Loupy et Helena Blancafort. Cette société française composée d'une douzaine de collaborateurs (des ingénieurs linguistes et informaticiens), propose des outils issus du traitement automatique du langage à destination des médias et d'acteurs du e-commerce et du e-tourisme.

La société a notamment développé, en partenariat avec les Echos, un agrégateur de flux d'information automatisé Les Echos 360, en juillet 2013. « C'est un "agrefilter", un agrégateur de flux qui analyse plus de 300 sources d'information économique originales (médias, blogs...), et qui les restitue en direct, de manière classée sur une seule page, en fonction de leur poids réel dans l'actualité et du traitement qui en est fait dans les médias », indiquait Frédéric Filloux, directeur des nouveaux médias des Echos, interrogé par Satellinet au moment de son lancement (lire également Satellinet n°186). Plus récemment, **Syllabs a également travaillé avec Slate pour le lancement de Slate Reader,** agrégateur d'informations issues des réseaux sociaux (lire également Satellinet n°215), pour lequel il a mis en place un système de tagging automatisé des contenus.

Outre la curation et l'agrégation de contenus, Syllabs développe des algorithmes destinés à la valorisation d'archives. La start-up a ainsi travaillé à la mise en place de la web-application « Un livre un jour » pour France 3 , réalisée avec WeDoData, agence positionnée sur le datajournalisme, et France Télévisions Nouvelles Ecritures. Elle permet d'accéder à toutes les archives de l'émission en fonction de divers critères (genre, lieu, époque...).

D'autres projets, plus ponctuels, ont également été menés pour différents médias dans le domaine de la datavisualisation, notamment la plateforme « Statsn'tweets », en partenariat avec Eurosport et TF1 au moment de l'Euro 2012, permettant de comparer les tweets générés sur des joueurs avec les données de la société Opta, et une web-application de comparaison des discours franco-allemands pour RadioFrance, en partenariat avec WeDoData.

Dans le domaine de la génération de texte, Syllabs a lancé la marque dédiée data2content, dont les clients actuels sont issus de l'e-tourisme, de l'e-commerce, des annuaires et des comparateurs en ligne, afin notamment d'optimiser leur SEO. « Les médias sont très demandeurs de ce type de technologie et nous sommes en discussions avancées avec plusieurs groupes pour le déploiement de projets en 2015. Data2content était jusqu'ici opérée comme une marque de Syllabs, des réflexions sont en cours pour en faire une société à part entière, tandis que Syllabs va se concentrer sur les projets Média et sémantique », indique Claude de Loupy, cofondateur de Syllabs.

« Les médias sont très demandeurs de ce type de technologie et nous sommes en discussions avancées avec plusieurs groupes pour le déploiement de projets en 2015. »

Claude de Loupy (Syllabs)

>>>

LE TOUR DE LA QUESTION

>>>

MELTYGROUP ET TRENDSBOARD RESTENT SUR LA DÉTECTION DES TENDANCES

Les acteurs français qui utilisent les algorithmes dans une optique de recommandation éditoriale, semblent, eux, plutôt en retrait sur la génération automatisée d'articles. A l'image de meltygroup, qui utilise son algorithme interne « Shape » afin d'identifier les sujets les plus discutés en temps réel sur les réseaux sociaux, les blogs et les sites médias. La production d'articles automatisés via cet outil n'est pas envisagée.

« L'axe de développement pour Shape concerne plutôt le timing des prévisions, afin de pouvoir anticiper les tendances plus en amont, mais aucunement la production éditoriale, qui reste l'apanage de nos rédacteurs. Étant orienté vers le divertissement pour les jeunes, l'automatisation de l'écriture n'aurait pas vraiment de sens », estime Olivier Levard, nouveau directeur des rédactions de meltygroup, par ailleurs auteur de l'ouvrage « Nous sommes tous des robots ».

Pour Trendsboard, plateforme Saas de recommandation éditoriale à destination des marques et des médias, la génération de contenus ne fait pas partie non plus des priorités, même si elle n'est pas écartée. « Étant positionnés sur le big data et la sémantique, il s'agit évidemment d'une piste de développement intéressante, et nous discutons avec plusieurs de nos partenaires à ce sujet. Il faut en revanche être sûr qu'il existe un marché », indique Benoît Raphaël, fondateur et CEO de Trendsboard, qui comptent parmi ses clients Radio France, Le Point, L'Express et Metronews.

CERTAINS OUTILS S'ORIENTENT VERS LE BTOC

Alors que la plupart des projets français et américains mis en place s'inscrivent dans une modèle BtoB (pour les médias, les entreprises, les marques....), dautres acteurs ont adopté un modèle davantage tourné vers le BtoC, à l'image de **Trooclick**, start-up française basée à Lyon et fondée en 2012. Celle-ci **propose une application de vérification des faits et de signalement des erreurs sur des articles de presse économiques et financiers**. Cette application, disponible seulement pour des contenus anglais, peut constituer un outil à la fois pour les journalistes et pour les internautes, notamment les professionnels de la finance, qui seraient prêts à payer pour ce type de service.

L'application, en version bêta, se présente sous la forme d'une extension pour navigateur Firefox, qui analyse le texte et interroge automatiquement des bases de données officielles afin d'identifier d'éventuelles erreurs, dans le cas d'introduction en bourse par exemple. Ce système fonctionnerait déjà pour les rubriques financières de journaux comme le New York Times, le Wall Street Journal, et Guardian, et pour les agences de presse Reuters et Bloomberg. Aucune déclinaison ne serait prévue en français dans les années à venir selon Stanislas Motte, CEO de Trooclick, interrogé par Le Monde.

Toujours dans le domaine de la vérification des faits, aux Etats-Unis, **le Washington Post expérimente depuis 2013, une application permettant une retranscription des discours politiques et leur croisement automatique avec les bases de données du journal**, afin de vérifier la véracité et la précision ou non des chiffres et des arguments fournis par les orateurs. Encore au stade de prototype, ce projet, baptisé « Truthteller » , repose sur un partenariat avec Microsoft pour la retranscription, et avec d'autres éditeurs comme The Texas Tribune pour les bases de données

A l'inverse, d'autres éditeurs ont choisi de développer des algorithmes permettant de transformer automatiquement des articles en flux audio. Le groupe Tribune Company a notamment lancé en 2014 l'application « Newsbeat », permettant de retranscrire, via un dispositif text-to-speech automatisé de convertir plusieurs milliers d'articles par jour en format audio, issus dela plupart des titres de presse américains via leur flux RSS**.** L'utilisateur indique ses centres d'intérêt et les sources qu'il souhaite agréger, et l'application, disponible sur iOS et Android. Une version gratuite avec publicité est proposée, ainsi qu'une formule premium à 1,99 dollar par mois ou 19,99 par an.

Fondée en 2010 à Tel Aviv, Wibbitz va plus loin en proposant des résumés vidéo d'articles de presse. L'application, disponible sous iOS uniquement, récupère automatiquement, via un partenariat avec Reuters et Getty News, des images, des vidéos et des infographies pour illustrer l'article, et génére un commentaire audio. Wibbitz aurait déja conclu des partenariats avec des dizaines de médias, principalement israéliens et américains (The Telegraph, Forbes Magazine, et le Daily Mirror, TechCrunch, etc.). Le modèle économique repose sur des publicités intégrées au sein des vidéos, dont les revenus sont partagés avec les éditeurs partenaires. Wibbitz propose par ailleurs différents formats aux éditeurs d'intégrer ces vidéos sur leur propre site.

Outre l'augmentation de la production éditoriale à moindre coût, les technologies de traitement automatique du langage et de la parole peuvent constituer de nouveaux débouchés en termes de produits pour les éditeurs, dans une approche davantage tournée vers le BtoC. ■