APERÇU

Application d'algorithmes de traitement du langage naturel pour extraire des informations d'articles de presse dans le cadre de la surveillance événementielle

Victoria Ng[1]*, Erin E. Rees[1], Jingcheng Niu[2], Abdelhamid Zaghlool[3], Homeira Ghiasbeglou[3], Adrian Verster[4]

Résumé

Cet article porte sur l'application du traitement du langage naturel (TLN) pour l'extraction d'informations dans les systèmes de surveillance événementielle (SSE). Nous décrivons les applications courantes de l'extraction d'informations à partir d'articles de presse et de sources médiatiques de sources ouvertes dans les SSE, les méthodes, la valeur en matière de santé publique, les difficultés et les nouveaux développements.

Citation proposée : Ng V, Rees EE, Niu J, Zaghlool A, Ghiasbeglou H, Verster A. Application d'algorithmes de traitement du langage naturel pour extraire des informations d'articles de presse dans le cadre de la surveillance événementielle. Relevé des maladies transmissibles au Canada 2020;46(6):211–7.
https://doi.org/10.14745/ccdr.v46i06a06f

Mots-clés : traitement du langage naturel, TLN, surveillance événementielle, algorithmes, extraction d'informations, données de sources ouvertes

Affiliations

[1] Laboratoire national de microbiologie, Agence de la santé publique du Canada

[2] Département de sciences informatiques, Université de Toronto, Toronto, ON

[3] Centre de mesures et d'interventions d'urgence, Agence de la santé publique du Canada

[4] Direction des aliments, Santé Canada, Ottawa, ON

***Correspondance :**
victoria.ng@canada.ca

Contexte

Les méthodes de traitement du langage naturel (TLN) permettent aux ordinateurs d'analyser, de traiter et de tirer un sens du discours humain. Le domaine du TLN existe depuis les années 1950 (1); toutefois, les progrès de la technologie et des méthodes de ces dernières années ont rendu les applications de TLN plus faciles à mettre en œuvre, certaines tâches menant à des meilleurs résultats que les performances humaines (2). Il existe de nombreuses applications quotidiennes du TLN, notamment la traduction automatique, la reconnaissance du pourriel et la reconnaissance vocale. Le TLN est un outil puissant dans le domaine des soins de santé en raison des volumes importants de données textuelles qui sont produites, par exemple les dossiers de santé électroniques. En effet, les dossiers de santé électroniques ont déjà fait l'objet d'applications du TLN, notamment pour la détection des proliférations mélanocytaires (3,4), du risque de démence (5) et des phénotypes neurologiques (6). Cependant, les applications du TLN dans le domaine des soins de santé vont au-delà des dossiers médicaux électroniques. Par exemple, il est possible d'identifier les personnes atteintes de la maladie d'Alzheimer à partir de leurs habitudes langagières (7).

Cet article porte principalement sur l'application du TLN pour l'extraction d'informations dans les systèmes de surveillance événementielle (SSE). Nous décrivons les applications courantes de l'extraction d'informations à partir d'articles de presse et de sources médiatiques de sources ouvertes dans les SSE, les méthodes, la valeur en matière de santé publique, les difficultés et les nouveaux développements.

Les SSE exploitent l'internet pour trouver des données de sources ouvertes, en s'appuyant sur des sources informelles (e.g. les activités des médias sociaux) et des sources formelles (e.g. les rapports médiatiques ou épidémiologiques des individus, des médias ou des organisations de santé) pour aider à détecter les menaces émergentes (8). Les systèmes opérationnels comprennent le Réseau mondial de renseignement de santé publique (RMISP) de l'Agence de la santé publique du Canada (9), HealthMap (10) et le renseignement sur les épidémies provenant de sources ouvertes de l'Organisation mondiale de la Santé (11). En raison de la variété, de la rapidité et du volume croissants des informations numériques, une multitude de données non structurées à source ouverte

sont générées quotidiennement, principalement sous forme de communications verbales ou écrites (9). Les données de sources ouvertes non structurées contiennent des informations pertinentes sur les menaces émergentes qui peuvent être traitées pour extraire des données structurées du bruit de fond afin de faciliter la détection précoce des menaces (12). Pour les SSE, cela comprend des informations sur la nature de l'événement (classification de la menace; nombre de cas), le lieu de l'événement (géolocalisation) et le moment de l'événement (informations temporelles). La capacité à identifier ces informations permet aux gouvernements et aux chercheurs de surveiller les menaces de maladies infectieuses émergentes et d'y répondre.

L'un des défis de la surveillance des maladies infectieuses, comme la COVID-19, est qu'une immense quantité de données textuelles est continuellement générée et, dans une pandémie en cours, cette quantité peut être bien supérieure à ce que les humains sont capables de traiter. Les algorithmes du TLN peuvent aider à ces efforts en automatisant le filtrage de grands volumes de données textuelles afin de trier les articles par degré d'importance et d'identifier et d'extraire les éléments d'information importants.

Dans cet article, nous discutons de certains algorithmes importants de TLN et de la manière dont ils peuvent être appliqués à la santé publique. Consultez le **tableau 1** pour un glossaire de la terminologie technique courante en TLN.

Tableau 1 : Glossaire de la terminologie technique courante dans le domaine du traitement du langage naturel

Terme	Définition
Annotation (linguistique)	L'association de notations descriptives ou analytiques avec des données linguistiques, généralement effectuée pour générer un corpus pour l'entraînement des algorithmes
Apprentissage automatique (AA)	Étude des algorithmes informatisés qui apprennent des modèles à partir d'expériences. Les approches de l'AA comprennent l'apprentissage supervisé (l'algorithme apprend à partir d'échantillons d'apprentissage étiquetés), non supervisé (l'algorithme créé des modèles à partir de données non étiquetées), ou semi-supervisé (l'algorithme apprend à l'aide d'une petite quantité de données étiquetées et une grande quantité de données non étiquetées)
Apprentissage non supervisé	Un type de méthode d'AA qui n'utilise pas données étiquetées, mais plutôt des approches analytiques de regroupement et de composantes principales afin que l'algorithme puisse trouver des attributs communs pour regrouper les données en différents résultats
Apprentissage supervisé	Les algorithmes issus de l'apprentissage supervisé représentent un type d'algorithme de l'AA qui apprend à partir de paires entrée-sortie étiquetées. Les attributs des données d'entrée sont extraits automatiquement grâce à l'apprentissage, et les modèles sont généralisés à partir de ces attributs afin de produire des prédictions sur les sorties. Les algorithmes

Tableau 1 : Glossaire de la terminologie technique courante dans le domaine du traitement du langage naturel (*suite*)

Terme	Définition
Apprentissage supervisé (*suite*)	communs incluent les modèles cachés de Markov (HMM), les arbres décisionnels, les modèles d'estimation basés sur l'entropie maximale, les machines à vecteur de support (SVM) et les champs conditionnels aléatoires (CRF)
Corpora (singulier - corpus)	Un ensemble d'articles où le texte non structuré a été annoté (étiqueté) pour identifier différents types d'entités nommées. Des corpora sont conçus pour différents domaines afin d'entraîner les algorithmes d'AA à identifier des entités nommées (e.g. le corpus WikToR d'articles Wikipédia pour les emplacements géographiques, le corpus TimeBank de nouveaux documents de rapport pour les informations temporelles)
Entité nommée (EN)	Un mot ou une phrase qui identifie un élément ayant des attributs particuliers qui le distinguent d'autres éléments ayant des attributs similaires (e.g. une personne, une organisation, un lieu)
Fils RSS	RSS signifie Really Simple Syndication ou Rich Site Summary. Il s'agit d'un type de flux Web qui permet aux utilisateurs et aux applications de recevoir des mises à jour régulières et automatisées à partir du site Web de leur choix sans avoir à visiter manuellement ces sites Web pour obtenir les mises à jour
Géocodage	Également appelé géorésolution, il attribue des coordonnées géographiques à des toponymes
Géolocalisation	Un sous-ensemble de la REN qui permet d'identifier des entités géographiques dans un texte non structuré
Géoparsing	Le processus combiné du géomarquage et du géocodage
Intelligence artificielle (IA)	Une branche de l'informatique traitant de la simulation de l'intelligence humaine par des machines
Linguistique computationnelle (LC)	Branche de l'informatique qui tente de modéliser le langage humain (y compris divers phénomènes linguistiques et applications liés au langage) à l'aide d'algorithmes computationnels
Précision (également appelée valeur prédictive positive)	Pourcentage des entités nommées trouvées par l'algorithme qui sont correctes : (vrais positifs) / (vrais positifs + faux positifs)
Polysémie	L'association d'un mot ou d'une phrase ayant deux ou plusieurs significations distinctes (e.g. une souris est un petit rongeur ou un dispositif de pointage pour un ordinateur)
Rappel	Fraction du montant total des cas pertinents qui ont été effectivement récupérés (vrais positifs) / (vrais positifs + faux négatifs)
Reconnaissance d'entité nommée (REN)	Le processus d'identification d'un mot ou d'une phrase qui représente une EN dans le texte. La REN est anciennement apparue dans la sixième conférence de compréhension des messages (MUC-6), à partir de laquelle les REN ont été classés en trois catégories : ENAMEX (personne, organisation, lieu), TIMEX (date, heure) et NUMEX (argent, pourcentage, quantité)

Tableau 1 : Glossaire de la terminologie technique courante dans le domaine du traitement du langage naturel (*suite*)

Terme	Définition
Score F1	Une mesure de la performance utilisée pour évaluer la capacité du TLN à identifier correctement les EN en calculant la moyenne harmonique de précision et de rappel : F1 = 2 * Précision * Rappel / (Rappel + Précision). Le score F1 privilégie les algorithmes équilibrés, car il tend vers le nombre le plus faible, minimisant l'incidence des grandes valeurs aberrantes et maximisant l'incidence des petites valeurs
Semi-supervisé	En raison des coûts élevés requis pour la création de données annotées, les algorithmes issus de l'apprentissage semi-supervisé apportent un équilibre coût-performance grâce à un apprentissage combiné à partir d'une petite quantité de données étiquetées (supervisé) et une grande quantité de données non étiquetées (non supervisé)
Synonymes	Les mots d'une même langue qui ont la même signification ou presque
Toponyme	L'EN du nom d'un lieu géographique tel qu'un pays, une province et une ville
Traitement du langage naturel (TLN)	Un sous-domaine de l'IA pour traiter les entrées en langage humain (naturel) pour diverses applications, y compris la reconnaissance automatique de la parole, la compréhension du langage naturel, la génération du langage naturel et la traduction automatique

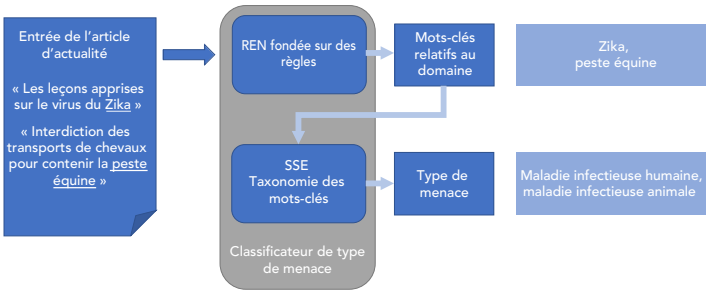
Les algorithmes de TLN et leur application à la santé publique

Le moyen le plus simple d'extraire des informations à partir de données textuelles non structurées est la recherche par mot-clé. Bien qu'efficace, cette méthode omet la question des synonymes et des concepts connexes (e.g. les nausées et vomissements sont liés aux maladies de l'estomac); elle ignore également le contexte de la phrase (e.g. en anglais, Apple peut être soit le fruit ou l'entreprise). Le problème de l'identification et de la classification des mots importants (entités) en fonction de la structure de la phrase est connu sous le nom de reconnaissance des entités nommées (REN) (13). Les entités les plus courantes sont les personnes, les organisations et les lieux. De nombreuses méthodes de REN étaient fondées sur des règles, identifiant et classant les mots à l'aide de dictionnaires (e.g. dictionnaire des noms d'agents pathogènes) et de règles (e.g. en utilisant « H#N# » pour classer une nouvelle souche de grippe non trouvée dans le dictionnaire) (14). Les synonymes et les concepts connexes peuvent être résolus à l'aide de bases de données qui organisent la structure des mots dans la langue (e.g. WordNet (15)). Les méthodes de REN les plus récentes utilisent des classifications et des relations prédéfinies dans les corpora pour développer des algorithmes d'apprentissage automatique (AA) afin d'identifier et de classifier les entités (13). Aux fins de la REN, les termes sont annotés en catégories et l'algorithme apprend à reconnaître d'autres exemples de la catégorie à partir du terme et de la structure de la phrase qui l'entoure. Comme les données linguistiques sont converties en jetons de mots dans le cadre de l'analyse, les algorithmes de TLN ne se limitent pas aux langues utilisant l'alphabet latin; ils peuvent également être utilisés avec des langues à base de caractères comme le chinois.

1. Classification des articles (type de menace)

La classification des articles par mots-clés taxonomiques en types de menaces permet aux utilisateurs du SSE de hiérarchiser les menaces émergentes. Par exemple, les analystes qui surveillent un événement peuvent filtrer les articles pour se concentrer sur une catégorie de menace spécifique. La REN fondée sur des règles identifie des mots-clés afin de relier chaque article à différentes catégories de menaces pour la santé (e.g. le type de maladie). Les mots-clés sont ensuite organisés en une taxonomie multilingue prédéterminée (e.g. « virus Zika » est une maladie infectieuse humaine, « peste équine » est une maladie infectieuse animale, etc.) qui peut être mise à jour à mesure que de nouvelles menaces sont découvertes. La taxonomie tire profit d'une structure linguistique similaire à celle de WordNet (16). Cela permet d'atténuer une partie du problème de la correspondance des mots-clés, car elle permet aux synonymes et aux concepts connexes de se substituer les uns aux autres. (**figure 1**).

Figure 1 : Classification des articles



Abréviations : REN, reconnaissance d'entités nommées; SSE, surveillance événementielle

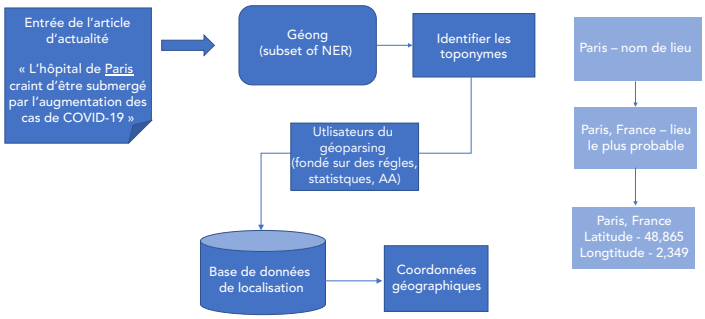
2. Géoparsing

L'identification des lieux où des événements liés à la santé sont rapportés à partir d'articles peut aider à localiser les populations sensibles. Le géoparsing consiste à attribuer des coordonnées géographiques à des entités de localisation (c'est-à-dire des toponymes tels que la ville, le pays) identifiées dans un texte non structuré. Le processus commence par la géolocalisation, un sous-ensemble de REN pour identifier les toponymes, puis se poursuit avec le géocodage pour attribuer des coordonnées géographiques à partir d'un dictionnaire comme celui des noms géographiques (17). Les utilisateurs du géoparsing font appel à des méthodes de calcul qui sont fondées sur des règles, des statistiques et l'AA. L'approche générale du géoparsing consiste à caractériser les toponymes par un ensemble d'éléments (e.g. le nom du toponyme, la position du premier et du dernier caractère dans le texte, la longueur en caractères). Les informations sur les caractéristiques sont ensuite traitées au moyen de méthodes computationnelles pour relier chaque toponyme à un

nom géographique dans une base de données de localisation (e.g. GeoNames (17)), puis lui attribuer les coordonnées correspondantes (18).

Les avancées en géoparsing, comme d'autres applications du TLN, visent à accroître la force des textes non structurés pour résoudre les ambiguïtés. Un progrès est l'utilisation de techniques d'apprentissage semi-supervisées qui utilisent des corpora générés par des programmes pour entraîner des algorithmes d'AA à partir de plus grands ensembles de données d'exemples annotés. L'utilisation de code pour annoter des articles est plus rapide et permet d'obtenir des corpora plus grands et plus cohérents que l'annotation humaine (19). L'exploitation d'un contexte plus large résulte également de l'élargissement des informations sur les caractéristiques pour qu'elles soient topologiques (relations spatiales entre les toponymes, e.g. la distance par rapport au toponyme voisin le plus proche) (20). Un toponyme tiré d'une phrase comme « il y a de nouveaux cas de grippe à Londres » peut être difficile à résoudre, car il existe de multiples lieux potentiels. Les coordonnées toponymiques peuvent être résolues en attribuant un biais en faveur des zones plus peuplées, car elles sont généralement mentionnées plus souvent dans le discours; cependant, les maladies émergentes ne favorisent pas toujours les zones très peuplées (**figure 2**).

Figure 2 : Géoparsing



Abréviations : AA, apprentissage automatique; REN, reconnaissance d'entités nommées

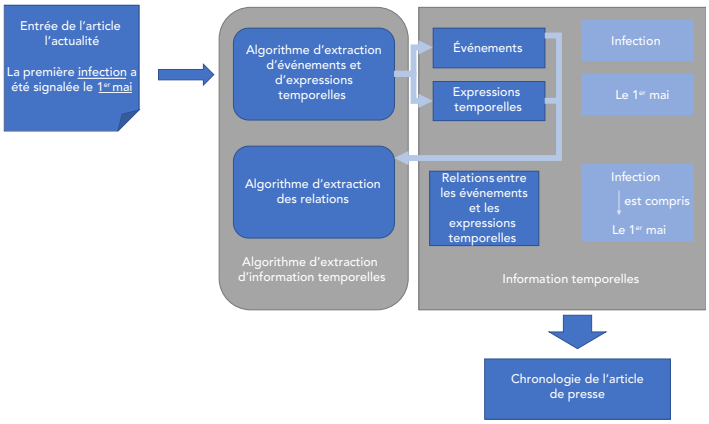
3. Extraction d'informations temporelles et raisonnement temporel

L'identification du moment où se produisent les événements décrits dans les articles est nécessaire pour établir une chronologie cohérente de ces événements. Il est important de pouvoir différencier un article rapportant un nouvel événement d'un article rapportant un événement antérieur connu. Les identificateurs temporels les plus courants dans les SSE sont la date de publication de l'article et la date de réception ou d'importation (l'horodatage de la réception de l'article dans le SSE). Aucune de ces dates n'extrait le moment des événements décrits dans les articles. Un sous-ensemble du TLN — l'extraction d'informations temporelles — a été développé pour extraire ces informations. L'extraction d'informations temporelles est utilisée pour identifier les jetons dans le texte qui contiennent des informations temporelles sur des événements pertinents.

Deux sous-tâches d'extraction d'informations temporelles aident à résoudre les ambiguïtés découlant de récits compliqués relatant des événements multiples. Premièrement, l'extraction de relations temporelles se concentre sur la classification des relations temporelles entre les événements extraits et les expressions temporelles. En utilisant ces relations, les SSE peuvent ancrer les événements dans le temps (e.g. dans la phrase « la première infection a été signalée le 1er mai », la relation entre l'événement « infection » et la date « 1er mai » est utilisée pour horodater la première infection). Deuxièmement, le raisonnement temporel (21) se concentre sur l'ordonnancement chronologique des événements par inférence.

Plusieurs systèmes d'extraction d'informations temporelles ont été développés, notamment TimeML (développé pour l'extraction temporelle d'articles de presse dans le domaine de la finance) (22); ISO-TimeML (une version révisée de TimeML) (23); et THYME (développé pour l'extraction temporelle dans les dossiers des patients) (24). Les résultats ont démontré l'atteinte d'une performance quasi humaine (25–28). En se fondant sur ces normes d'annotation, une norme d'annotation pour les articles de presse dans le domaine de la santé publique, Temporal Histories of Epidemic Events (THEE), a récemment été développée pour les SSE par les auteurs de cet article (29) (**figure 3**).

Figure 3 : Extraction d'informations temporelles et raisonnement temporel



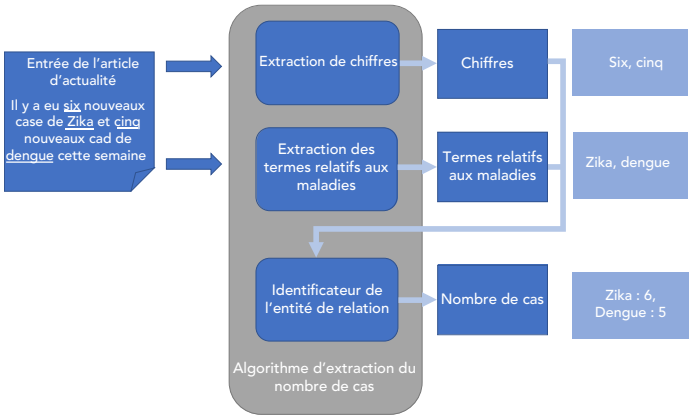
4. Extraction du nombre de cas

L'extraction du nombre de cas de maladie signalés dans les articles aiderait les utilisateurs d'un SSE à surveiller et à prévoir la progression de la maladie. Actuellement, il n'existe pas d'algorithme de TLN intégré dans les SSE capable de cette tâche. Cependant, il existe des algorithmes capables de s'attaquer à des tâches connexes qui peuvent être exploités pour développer un algorithme de comptage des cas. Les articles de presse en épidémiologie mentionnent fréquemment l'apparition

de cas de maladie (e.g. « Il y a eu six nouveaux cas de Zika cette semaine »), de sorte que l'identification des cas nécessite de déterminer les relations entre une référence quantitative dans le texte (six nouveaux cas) et un terme de maladie (de Zika). De nombreux algorithmes identifient déjà les relations entre des entités dans divers domaines. Par exemple, l'algorithme RelEx identifie les relations entre les gènes qui sont enregistrés dans les résumés MEDLINE et fonctionne avec un score F1 de 0,80 (30). Un algorithme a été développé à partir de l'algorithme RelEx, pour identifier les phrases dans les articles de presse qui font état de cas de maladies d'origine alimentaire (31).

Les auteurs de cet article développent et affinent cet algorithme pour extraire des informations sur le nombre de cas à partir de phrases qui ont été identifiés comme contenant des informations sur le nombre de cas (**figure 4**).

Figure 4 : Extraction du nombre de cas

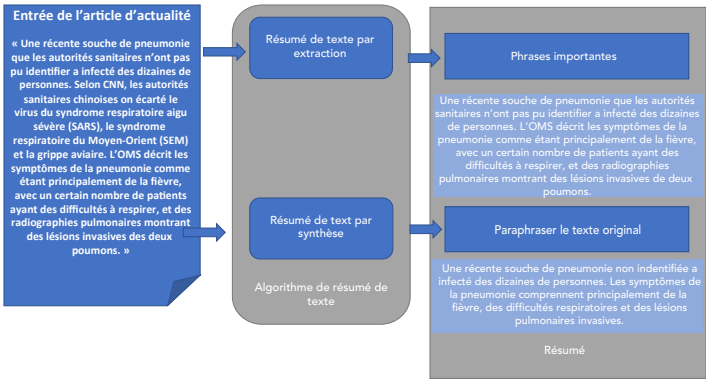


5. Résumé automatique du texte

L'objectif de la synthèse de texte est de créer rapidement et précisément un résumé concis qui conserve les informations essentielles du texte original. Le résumé de texte dans les SSE augmenterait le nombre d'articles qui peuvent être parcourus pour y détecter des menaces en réduisant le volume de texte à lire. Il existe deux principaux types de résumés de texte : soit un résumé fondé sur l'extraction et l'autre fondé sur la synthèse. La synthèse par extraction consiste à identifier les mots et phrases clés les plus importants du texte et à les combiner mot à mot pour produire un résumé. Le résumé fondé sur la synthèse utilise une technique plus sophistiquée qui consiste à paraphraser le texte original pour écrire un nouveau texte, imitant ainsi le résumé d'un texte humain.

Le résumé de texte en TLN est normalement développé à l'aide de modèles d'AA supervisés et formés sur des corpora. Dans les deux cas, soit le résumé par extraction et le résumé par synthèse, les phrases principales sont extraites du document source à l'aide de méthodes telles que le balisage des parties de discours, les séquences de mots ou d'autres méthodes de reconnaissance des formes linguistiques (32). Le résumé par synthèse va plus loin et tente de créer de nouvelles phrases et expressions à partir des phrases principales extraites. Un certain nombre de techniques sont utilisées pour améliorer le degré de synthèse, notamment des techniques d'apprentissage approfondi et des modèles de langue préformés (33) (**figure 5**).

Figure 5 : Résumé automatique du texte



Abréviation : OMS, Organisation mondiale de la Santé

Discussion

Le domaine des soins de santé offre un nombre énorme d'applications potentielles pour le TLN en raison de l'omniprésence des données textuelles. Les dossiers de santé électroniques sont une source évidente de données pour l'application du TLN, mais les textes relatifs aux soins de santé vont bien au-delà des dossiers de santé; ils incluent les sources traditionnelles et les médias sociaux, qui sont les principales sources de données pour les SSE, en plus des rapports et documents officiels des gouvernements.

Comme les algorithmes de TLN peuvent interpréter des textes et extraire des informations essentielles de sources de données aussi diverses, ils continueront à jouer un rôle croissant dans la surveillance et la détection des maladies infectieuses émergentes. L'actuelle pandémie de COVID-19 est un exemple de cas où les algorithmes de TLN pourraient être utilisés pour la surveillance des crises de santé publique. (C'est d'ailleurs ce que plusieurs coauteurs de cet article sont en train de développer.)

Toutefois, même s'ils sont puissants, les algorithmes du TLN ne sont pas parfaits. Les principaux défis actuels consistent à regrouper plusieurs sources faisant référence à un même événement et à traiter les imperfections dans la précision de l'extraction des informations dues aux nuances des langues humaines. Les recherches en TLN sur l'extraction d'informations de la prochaine génération qui peuvent améliorer ces défis comprennent la résolution d'événements (déduplication et liaison des mêmes événements entre eux) (34) et les avancées dans les approches de TLN neurales, comme les réseaux de transformateurs (35), le mécanisme d'attention (36) et les modèles de langage à grande échelle, comme ELMo (37), BERT (38) et XLNet (39), visant à améliorer les performances actuelles des algorithmes.

Conclusion

Nous avons discuté de plusieurs algorithmes d'extraction de TLN communs aux SSE : la classification des articles, qui peut identifier les articles contenant des informations cruciales sur la propagation des maladies infectieuses; la géolocalisation, qui identifie où un nouveau cas de la maladie s'est produit; l'extraction temporelle, qui identifie quand un nouveau cas s'est produit; l'extraction du nombre de cas, qui identifie combien de cas se sont produits; et le résumé des articles, qui peut réduire considérablement la quantité de texte à lire par un humain.

Bien que le domaine du TLN pour l'extraction d'informations soit bien établi, de nombreux développements existants et émergents pertinents pour la surveillance de la santé publique se profilent à l'horizon. S'ils sont mis à profit, ces développements pourraient se traduire par une détection plus précoce des nouvelles menaces sanitaires ayant des répercussions immenses sur les Canadiens et le monde entier.

Conflit d'intérêts

Aucun.

Financement

E. E. Rees et V. Ng sont actuellement les co-enquêteurs principaux du Programme canadien pour la sûreté et la sécurité (PCSS), un programme financé par le ministère de la Défense nationale. La subvention est une subvention de trois ans intitulée « Incorporating Advanced Data Analytics into a Health Intelligence Surveillance System » — CSSP-2018-CP-2334.

Références

1. Bates M. Models of natural language understanding. Proc Natl Acad Sci USA 1995 Oct;92(22):9977–82. DOI PubMed

2. Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, Levy O, Bowman S. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. Ithaca (NY): arXiv; 2019 (mise à jour 2020-02-13; accédé 2020-02-24). https://arxiv.org/abs/1905.00537

3. Lott JP, Boudreau DM, Barnhill RL, Weinstock MA, Knopp E, Piepkorn MW, Elder DE, Knezevich SR, Baer A, Tosteson AN, Elmore JG. Population-based analysis of histologically confirmed melanocytic proliferations using natural language processing. JAMA Dermatol 2018 Jan;154(1):24–9. DOI PubMed

4. Nguyen AN, Truran D, Kemp M, Koopman B, Conlan D, O'Dwyer J, Zhang M, Karimi S, Hassanzadeh H, Lawley MJ, Green D. Computer-Assisted Diagnostic Coding: effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings. AMIA Annu Symp Proc 2018 Dec;2018:807–16. PubMed

5. McCoy TH Jr, Han L, Pellegrini AM, Tanzi RE, Berretta S, Perlis RH. Stratifying risk for dementia onset using large-scale electronic health record data: A retrospective cohort study. Alzheimers Dement 2020 Mar;16(3):531–40. DOI PubMed

6. Wheater E, Mair G, Sudlow C, Alex B, Grover C, Whiteley W. A validated natural language processing algorithm for brain imaging phenotypes from radiology reports in UK electronic health records. BMC Med Inform Decis Mak 2019 Sep;19(1):184. DOI PubMed

7. Karlekar S, Niu T, Bansal M. Detecting linguistic characteristics of Alzheimer's Dementia by interpreting neural models. In: Proceedings of NAACL-HLT 2018. New Orleans (LA). June 1–6, 2018.

8. World Health Organization. A guide to establishing event-based surveillance. Manila (PH): WHO Regional Office for the Western Pacific; 2008.

9. Dion M, AbdelMalik P, Mawudeku A. Les données massives et le Réseau mondial d'information en santé publique (RMISP). Relevé des maladies transmissible au Canada 2015;41(9):241–7. DOI

10. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J Am Med Inform Assoc 2008 Mar-Apr;15(2):150–7. DOI PubMed

11. World Health Organization. Epidemic Intelligence from Open Sources (EIOS): saving lives through early detection. Geneva (CH): World Health Organization; 2020 (accéd 2020-01-24). https://www.who.int/eios

12. Barboza P, Vaillant L, Mawudeku A, Nelson NP, Hartley DM, Madoff LC, Linge JP, Collier N, Brownstein JS, Yangarber R, Astagneau P; Early Alerting Reporting Project Of The Global Health Security Initiative. Evaluation of epidemic intelligence systems integrated in the early alerting and reporting project for the detection of A/H5N1 influenza events. PLoS One 2013;8(3):e57252. DOI PubMed

13. Nadeau D, Sekine S. A survey of named entity recognition and classification. In: Nadeau D, Sekine S, editors. Named entities: recognition, classification and use. Lingvisticae Investigationes 2007;30(1):3–26.

14. Sekine S, Nabota C. Definition, dictionaries and tagger for extended named entity hierarchy. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004). Lisbon (PT). 26–28 May 2004.

15. Princeton University. WordNet: an electronic lexical database. Princeton (NJ): Princeton University; 2010 (accédé 2020-01-24). https://wordnet.princeton.edu/

16. Miller GA. WordNet: a lexical database for English. Commun ACM 1995;38(11):39–41.

17. GeoNames [database]. 2020. https://www.geonames.org/

18. Santos J, Anastácio I, Martins B. Using machine learning methods for disambiguating place references in textual documents. GeoJournal 2015;80(3):375–92. DOI

19. Gritta M, Pilehvar MT, Limsopatham N, Collier N. What's missing in geographical parsing? Lang Resour Eval 2018;52(2):603–23. DOI PubMed

20. DeLozier G, Baldridge J, London L. Gazetteer-independent toponym resolution using geographic word profiles. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin (TX): AAAI Press; 2015. p. 2382–8.

21. Allen JF. Maintaining knowledge about temporal intervals. In: Weld DS, De Kleer J. Readings in qualitative reasoning about physical systems. 1990, Elsevier; pp. 361–72. DOI

22. Pustejovsky J, Ingria R, Sauri R, Castano JM, Littman J, Gaizauskas RJ, Setzer A, Katz G, Mani I. The specification language TimeML. In: Mani I, Pustejovsky J, Gaizauskas R, editors. The language of time: a reader. 2005. p. 545–58.

23. Pustejovsky J, Kiyong L, Bunt H, Romary L. ISO-TimeML: an international standard for semantic annotation. In: Proceeding from the International Conference on Language Resources and Evaluation 2010. La Valette (MT). 2010 May.

24. Styler WF 4th, Bethard S, Finan S, Palmer M, Pradhan S, de Groen PC, Erickson B, Miller T, Lin C, Savova G, Pustejovsky J. Temporal annotation in the clinical domain. Trans Assoc Comput Linguist 2014 Apr;2:143–54. DOI PubMed

25. Chambers N. Navytime: event and time ordering from raw text. In: Second Joint Conference on Lexical and Computational Semantics (*SEM). Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Annapolis (MD): Naval Academy; 2013.

26. Lee HJ, Xu H, Wang J, Zhang Y, Moon S, Xu J, Wu Y. UTHealth at SemEval-2016 Task 12: an end-to-end system for temporal information extraction from clinical notes. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). San Diego (CA): Association for Computational Linguistics; 2016. DOI

27. Strötgen J, Zell J, Gertz M. HeidelTime: tuning English and developing Spanish resources for TempEval-3. In: Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM). Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Atlanta (GA);2013.

28. Lin C, Miller T, Dligach D, Bethard S, Savova G. A BERT-based universal model for both within- and cross-sentence clinical temporal relation extraction. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis (MN);2019.

29. Niu J, Ng V, Penn G, Rees E. Temporal histories of epidemic events (THEE): a case study in temporal annotation for public health. In: Proceedings of the International Conference on Language Resources and Evaluation. Marseille (FR);2020.

30. Fundel K, Küffner R, Zimmer R. RelEx--relation extraction using dependency parse trees. Bioinformatics 2007 Feb;23(3):365–71. DOI PubMed

31. Nasheri N, Vester A, Petronella N. Foodborne viral outbreaks associated with frozen produce. Epidemiol Infect 2019 Oct;147:e291. DOI PubMed

32. Aries A, Eddine ZD, Hidouci WK. Automatic text summarization: what has been done and what has to be done. arXiv:1904.00688. https://arxiv.org/abs/1904.00688

33. Kryscinski W, Paulus R, Xiong C, Socher R. Improving abstraction in text summarization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels (BE): Association for Computational Linguistics; 2018. DOI

34. Petroni F, Raman N, Nugent T, Nourbakhsh A, Panic Z, Shah S, Leidner J. An extensive event extraction system with cross-media event resolution. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.

35. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser L, Polosukhim I. Attention is all you need. In: Proceedings from the 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach (CA);2017.

36. Liu B, Lane I. Attention-based recurrent neural network models for joint intent detection and slot filling. Proc Interspeech 2016;685–9. DOI

37. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) 2018. New Orleans (LA). DOI

38. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR 2018;abs/1810.04805. https://arxiv.org/abs/1810.04805

39. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q. XLNet: generalized autoregressive pretraining for language understanding. arXiv 2019;1906.08237. https://arxiv.org/abs/1906.08237