Criminologie, Forensique et Sécurité

www.revue-cfs.net | ISSN : 2813-6098

Organe officiel de l'association internationale des criminologues de langue française (AICLF)

Article théorique

Criminologie, Forensique et Sécurité

2023 | 1 (1) : 3349

https://doi.org/ 10.26034/la.cfs.2023.3349

La reconnaissance d'entités nommées : une démarche prometteuse pour la détection automatisée de liens dans les documents d'enquête policière

Maxime Bérubé[1], Francis Fortin[2] et Olivier Péloquin[3]

[1] PhD, Chaire de recherche UQTR en forensique numérique, Université du Québec à Trois-Rivières
[2] PhD, Ecole de criminologie, Université de Montréal
[3] MSc, Ecole de criminologie, Université de Montréal.

Contact : Maxime.Berube2@uqtr.ca

Résumé

L'information produite par nos activités numériques est en constante augmentation. Ce flux d'informations en continu se traduit aussi par un accroissement important du nombre de données à traiter dans le cadre d'activités de renseignement et d'enquêtes policières. Afin de faciliter ce traitement de données, de nouvelles techniques ayant recours à l'intelligence artificielle s'offrent aux personnels policiers afin d'automatiser une partie de leur travail. Dans ce contexte, le présent article propose une démarche en six étapes permettant le déploiement d'une démarche structurée et d'un modèle algorithmique de reconnaissance d'entités nommées, spécifiquement adaptée pour l'analyse de documents d'enquête policière. En mettant l'accent plus spécifiquement sur le traitement de dossier d'infractions pour fraude, la démarche méthodologique à entreprendre pour avoir recours efficacement à ces nouvelles technologies d'analyse y est donc décrite en détail. De plus, l'évolution du rôle de l'analyste en renseignement criminel, l'acteur étant au cœur de l'intégration de ce type d'innovations, y est également discutée, tout en soulignant la pertinence de la reconnaissance d'entités nommées en contexte d'enquête policière.

Mots clés

linguistique forensique, enquête, renseignement, traitement du langage naturel, reconnaissance d'entités nommées

Named-entity recognition: a promising approach for the automated detection of links in police investigation file

Abstract

The information produced by our digital activities is constantly increasing. This continuous flow of information also results in a significant increase in the amount of data to be processed by law enforcement agencies in the context of intelligence activities and investigations. To facilitate this data processing, new techniques using artificial intelligence are available to them in order to automate part of their work. In this context, this article suggests a six-step approach for the deployment of a structured approach and an algorithmic model of named-entity recognition, especially build for the analysis of police investigation files. By emphasising more specifically the processing of fraud cases, the methodological approach to be taken to make effective use of these new analysis technologies is described in detail. In addition, the evolution of the role of the criminal intelligence analyst, the actor being at the heart of the integration of this type of innovation, is also discussed, while emphasising the relevance of the named-entity recognition in the context of a police investigation.

Keywords

forensic linguistics, investigation, intelligence, natural language processing, named entity recognition

Citation : Bérubé, M., Fortin, F., et Péloquin O. (2023) La reconnaissance d'entités nommées : une démarche. prometteuse pour la détection automatisée de liens dans les documents d'enquête policière. *Criminologie, Forensique et Sécurité*, 1 (1) : 3349.

Introduction

En cette ère de transformations sociales et technologiques affectant entre autres la nature de la criminalité et sa régulation, le recours aux méthodologies automatisées de traitement des données s'avère un incontournable pour les organisations policières (McGuire et Holt, 2017; Rossy et al., 2018). En effet, la quantité d'informations et de renseignements devant être traitée en contexte d'enquête, et particulièrement ceux issus de traces numériques, ne cesse d'augmenter. Au cours des dernières années, plusieurs services de police ont d'ailleurs vu le nombre de signalements et de dossiers d'enquête pour des activités frauduleuses s'accentuer de manière exponentielle, et ce, particulièrement pendant la crise sanitaire liée à la COVID-19 (Plouffe, 2021). Or, le traitement de ces signalements représente un travail colossal pour les analystes en renseignement, ou pour tout autre acteur impliqué dans un processus d'enquête, de même que des coûts très importants pour les organisations. La portée internationale, la complexité des enquêtes, la quantité et la variété des documents accessibles ont également rendu désuètes les pratiques conventionnelles s'appuyant sur un travail manuel d'analyse des données. Il est simplement devenu impossible de pouvoir lire, annoter et classer manuellement tous les éléments pertinents à l'enquête. Les pratiques d'enquête doivent donc être adaptées aux nouvelles réalités de l'ère numérique et il est impératif que les méthodologies actuelles de collecte et de traitement du renseignement soient revues afin d'y inclure, au moins partiellement, une automatisation des procédés (De Pauw et al., 2011; McCue, 2014).

Ainsi, l'une des composantes essentielles de l'enquête comprend la recherche et l'analyse d'information issue de sources documentaires (Oatley et Ewart, 2011). En fonction des besoins des organisations policières, de multiples techniques en linguistique forensique et d'intelligence artificielle sont mobilisées à cette fin, notamment en ce qui a trait au traitement du langage naturel et du forage de texte (Hassani et al., 2016). Au cours des dernières années, des avancées importantes ont été observées dans le domaine du traitement du langage naturel et de l'apprentissage machine en général. Ces avancées, en plus de rendre la technologie plus efficace grâce à des modèles d'apprentissages, ont aussi permis la démocratisation de son utilisation. C'est dans ce contexte que la présente étude discute de l'application de techniques automatisée du traitement de langage naturel en enquête criminelle.

Avec l'objectif de contribuer au traitement, au classement et à la priorisation des dossiers en contexte d'enquête policière, la présente contribution vise à présenter dans quelles mesures ces techniques de reconnaissance d'entités nommées (REN) peuvent s'avérer prometteuses pour faciliter la conduite d'enquêtes lorsqu'une telle quantité massive de données est observée. Dans les sections qui suivent, l'évolution des enquêtes et le rôle de l'analyste en renseignement sont d'abord discutés, suivis d'une mise en contexte de la pertinence du recours aux nouvelles technologies pour faire face à la panoplie de documents non structurés en contexte d'enquête. Par la suite, la troisième section souligne la pertinence de la REN pour la découverte de liens dans ce même contexte. Enfin s'en suit la quatrième section présentant la démarche en six étapes proposées pour la construction d'un algorithme de REN à cette fin, ainsi que les limites inhérentes à celle-ci. Tout au long de ce travail, le rôle incontournable de l'analyste en renseignement dans une démarche d'enquête ayant recours à l'intelligence artificielle y est souligné.

L'évolution des enquêtes et du rôle de l'analyste en renseignement

L'évolution de la criminalité ainsi que la nouvelle traçabilité des activités humaines par l'usage des technologies de l'information, de même que les traces numériques qu'elles engendrent, implique une transformation des stratégies de gouvernance de la sécurité et d'application de la loi (Grossrieder et al., 2013). L'enquête policière et le renseignement constituent des activités qui sont particulièrement affectées par cette transformation, notamment parce que l'une des composantes essentielles de celles-ci comprend l'analyse de quantités massives de données enregistrées dans différentes banques d'informations (Oatley et Ewart, 2011). Au lieu d'examiner chaque information manuellement, il est maintenant nécessaire d'utiliser des outils et des logiciels d'analyse afin d'orienter les enquêtes et de formuler des recommandations (Banarescu, 2015). Dans ce contexte, une attention particulière doit notamment être portée sur la qualité des données qui sont collectées afin de développer une culture de recherche plus rigoureuse, plus efficace et surtout permettant un meilleur rendement des outils employés (O'Connor, 2021). De cette façon, le développement de l'analyse criminelle permet d'appréhender le phénomène criminel au-delà des limites imposées par les mutations récentes de la technologie (Grossrieder et al., 2013).

Afin qu'une telle analyse rigoureuse soit possible, les agences d'application de la loi doivent généralement recourir à l'expertise d'analystes en renseignement : ceux-ci étant au cœur des innovations technologiques et du traitement de l'information. En effet, ces derniers collaborent aux enquêtes en structurant les données et en les traitant afin de faire ressortir les éléments d'intérêt pour chacun des dossiers traités (Brun, 2018). De façon générale, leur rôle consiste à émettre des hypothèses qui seront par la suite affinées et testées progressivement à partir de différents outils et différentes sources tels que les rapports policiers ou encore des informations collectées auprès de la population (Baechler et al., 2020). Les analystes en renseignement offrent également aux gestionnaires des capacités supplémentaires quant à l'efficacité et la qualité des services de sécurité offerts, entre autres en leur fournissant de meilleures connaissances liées aux phénomènes criminels auxquels ils sont confrontés (Cofan et Baloi, 2017).

Le rôle de l'analyste en renseignement consiste entre autres à organiser chacune des informations issues de différentes banques de données afin de leur donner un sens à partir duquel les données seront mieux comprises et mieux interprétées (Deering et Corkill, 2017; O'Connor, 2021). Ainsi, les fonctions de l'analyste constituent un moyen d'organiser les informations en un tout significatif afin d'améliorer le développement d'hypothèses et la prédiction de la criminalité (Harper et Harris, 1975). Son rôle en tant que producteur d'informations est central dans l'orientation des ressources policières, d'une part, pour alimenter les réflexions précédent les prises de décisions et, d'autre part, pour améliorer les performances policières (Deering et Corkill, 2017; Keay et Kirby, 2018; Piza et Feng, 2017). Bref, optimiser le rôle de l'analyste en renseignement permet aux organisations policières de limiter leurs ressources, tout en maximisant la qualité de leur réponse face aux enjeux de la criminalité (Osborne, 2001).

Les analystes ont à traiter des sommes astronomiques de données : l'utilisation d'outil de synthèse, de classement et de visualisation de données revient souvent sous leur responsabilité. Parmi les crimes à complexité élevée et impliquant des quantités importantes de données à traiter, on retrouve notamment la fraude, un type d'infraction en transformation constante visant à tirer parti des vulnérabilités humaines et informatiques les plus actuelles, par exemple le paiement sans contact ou les guichets de cryptomonnaie (Westphal, 2008). L'enquête sur ce type d'infraction génère des quantités impressionnantes de documents textuels produits notamment par des ordonnances et des historiques de communication, des relevés de transaction, des déclarations de témoins, des rapports d'enquête, etc. Ainsi, la mise en place de processus d'analyses dédiés aux opérations complexes permet d'optimiser les pratiques d'enquête traitant en particulier des infractions de fraude (Banarescu, 2015). Il devient donc impératif d'adapter les méthodologies actuelles afin de faire face aux opérations frauduleuses qui génèrent une quantité importante d'informations (Hipgrave, 2013). Sans systèmes de gestion de l'information à la fine pointe de la technologie, les analyses ne cesseront d'être affectées par le volume considérable et toujours croissant de données d'enquête à traiter (Osborne, 2001). Ainsi, le recours à l'automatisation et à l'analyse assistée par ordinateur comme stratégie d'analyse faciliterait l'identification des risques frauduleux de même que l'identification des tendances. À partir de ces systèmes, l'analyste maximise son efficacité, notamment pour l'examen de relations et de similarités entre les dossiers d'enquêtes.

Structurer le non structuré en contexte d'enquête

Alors que d'énormes quantités de contenu textuel sont partagées chaque jour, la nécessité de développer des moyens d'extraire des connaissances de ce type de traces numériques s'avère d'une importance capitale pour les analystes à la recherche de liens communs entre les dossiers qu'ils ont à traiter. Les données textuelles, notamment celles de dossiers d'enquête policière, sont des expressions du langage naturel qui ne peuvent pas être comprises facilement par les ordinateurs à moins qu'ils ne puissent être représentés et caractérisés de manière appropriée (Milić-Frayling, 2005). Utiles dans de nombreux domaines de l'activité humaine, des techniques de forage de textes (text mining) ont donc été développées pour récupérer des connaissances à partir de bases de données textuelles (Feldman et Dagan, 1995). Comme l'exploration de données en général (data mining), ces techniques se concentrent sur la localisation et l'extraction de motifs intéressants et non triviaux, mais elles peuvent également être utilisées pour obtenir des connaissances à partir de documents textuels non structurés (Tan, 1999). Dans ce contexte, les résultats probants et utiles se distinguent par une combinaison de pertinence, de nouveauté et d'intérêt (Han et al., 2012).

L'exploration de texte implique généralement le traitement linguistique et statistique des documents. La linguistique forensique, et la science linguistique plus largement, a étudié les moyens de caractériser et d'expliquer les productions linguistiques, en particulier les conversations et les écrits (Milić-Frayling, 2005). L'étude des structures linguistiques par lesquelles nous communiquons a conduit à la création d'outils qui caractérisent le langage (ex. : voir l'initiative Wordnet ; Vossen (2002)) et agissent comme un ensemble de règles. Bien que l'exploration de texte puisse être effectuée par l'ordinateur sans intervention humaine, les chercheurs ont suggéré que l'intégration des connaissances d'un domaine spécifique puisse améliorer les méthodes d'analyse syntaxique ainsi que l'efficacité de l'apprentissage et de l'exploration et la qualité du modèle (Tan, 1999). L'exploration de texte est donc considérée comme un domaine multidisciplinaire, où les tâches comprennent la catégorisation, le regroupement, l'extraction de concepts et d'entités, la modélisation entité-relation, les taxonomies, l'analyse des sentiments et le résumé de documents (Feldman et Sanger, 2007).

En contexte enquête, il a été démontré empiriquement que l'intégration d'un processus automatisé d'analyse de texte engendre une réduction des ressources nécessaires pour le classement de documents policiers (Asharef et al., 2012; Bsoul et al., 2013), la mise en relation de dossiers d'enquête (Arulanandam et Savarimuthu, 2014; Chen, Chung, et al., 2003; Hassani et al., 2016), la recherche d'information dans les bases de données ou les documents d'enquête (Chen, Schroeder, et al., 2003; Gianola, 2020; Hauck et al., 2002), ainsi que la prédiction de tendances et de « hot spots » de la criminalité (Hassani et al., 2016). C'est dans ce contexte que la reconnaissance d'entités nommées (REN), aussi appelée Named-Entity Recognition (Chau et al., 2002; Chen et al., 2004; Hassani et al., 2016), prend tout son sens dans le traitement des données d'enquête, où l'extraction de connaissance à partir de données non structurées est aujourd'hui essentielle à l'efficacité du travail des analystes en renseignement. En effet, cette technique permet d'apporter une aide substantielle pour accomplir l'une des principales tâches de l'analyse criminelle qui implique de lier différentes entités d'intérêt entre elles et de découvrir des séries criminelles (Harper et Harris, 1975; Rossy, 2016).

La REN pour la découverte de liens dans les enquêtes

La notion d'entités réfère à tout objet matériel ou conceptuel dont il est mention dans une enquête (Gianola, 2020). En pratique, n'importe quel objet ou sujet peut donc constituer une entité d'intérêt, qu'il s'agisse de personnes, d'événements, d'informations, de traces, etc. (Rossy et al., 2019). Selon Merry (2000), l'analyse des données afin d'établir des liens constitue l'essence même de l'analyse opérationnelle. Elle consiste en la décomposition et l'interprétation des données qui sont représentées sous la forme d'entités et de relations (Harper et Harris, 1975). Plus précisément, il s'agit de mettre en évidence des relations telles que des similitudes, des régularités ou des correspondances entre les informations collectées (Batura, 2021; Rossy, 2011). À partir de méthodes et de techniques spécifiques, ce procédé permet donc d'identifier les liens et des tendances qui sont difficilement traçables dans les enquêtes policières régulières (Banarescu, 2015; Merry, 2000). En effet, l'analyse de réseaux complexes contenant des milliers d'entités et de relations en ne se basant que sur les techniques d'enquête traditionnelles s'avère quasi impossible. L'usage de systèmes adaptés à la technologie facilite, dans ce contexte, la détection de ces entités d'intérêt (Rossy, 2016). Il est possible d'y déterminer la présence ou non de liens entre les individus et de recueillir des informations qui permettront d'identifier et de clarifier la nature et la force de ces relations (Harper et Harris, 1975; Rossy, 2016). En effet, l'intérêt d'identifier des entités ne repose pas seulement sur l'extraction de ces dernières,

mais également sur la nature des rapports qu'elles entretiennent (Rossy, 2011). De nombreux autres avantages tactiques et stratégiques en découlent, comme une évaluation approfondie de la structure interne des groupes criminels et l'identification de pseudonymes difficilement détectables dans les enquêtes régulières (Berlusconi et al., 2016). En analysant de cette façon les relations et les associations entre les entités recueillies, l'analyse des liens peut également fournir des informations sur les motifs ayant poussé les criminels à agir et ainsi orienter les différentes pistes d'enquête (Piza et Feng, 2017; Schroeder et al., 2007).

L'identification de relations entre différents dossiers d'enquête peut donc se faire en recherchant des entités communes inscrites dans les documents constituant ces différents dossiers. Pour ce faire, la REN vise à identifier de manière automatisée des entités telles que des noms de personnes, des organisations, des lieux, ou encore toutes autres entités pouvant s'avérer d'intérêt pour l'enquête. Les résultats obtenus peuvent ensuite permettre de constater des relations qui, en raison de la quantité de documents à traiter, auraient pu échapper au travail manuel d'un analyste en renseignement (Chau et al., 2002; Hauck et al., 2002). Le recours aux analyses par REN peut aussi permettre une première analyse automatisée de dossiers d'enquête complexes afin d'orienter rapidement les démarches d'enquête, et ce, avant même que les documents aient pu être analysés manuellement de manière conventionnelle. Par exemple, afin de soutenir la conduite d'entrevue d'enquête, Ku et al. (2008) ont eu recours à la REN pour extraire rapidement des informations pertinentes dans des déclarations de témoins. De cette façon, en fonction de la reconnaissance de certaines entités spécifiques et d'indices basés sur des techniques d'entrevue cognitive, les questions d'entrevues pouvaient être réorientées afin d'obtenir plus d'informations pertinentes des personnes rencontrées.

Plusieurs études ont évalué la pertinence de l'utilisation de la REN dans un contexte d'enquête. Toutefois, les outils développés jusqu'ici sont le plus souvent construits et validés sur des documents issus d'articles de presse ou de documents policiers rendus publics et ne sont donc pas représentatifs du format et des paramètres relatifs aux documents traités par les analystes au quotidien (Hassani et al., 2016). Dès lors, leur impact sur la gestion subséquente des ressources et de la criminalité demeure limité. Dans un même ordre d'idée, peu d'études en ce sens ont réellement permis de mettre à profit les techniques modernes d'intelligence artificielle. Par exemple, Gianola (2021) propose une approche uniquement basée sur une reconnaissance de structures linguistiques et d'entités inscrites dans des lexiques. Contrairement aux algorithmes d'intelligence artificielle, ce type de reconnaissance est tributaire dudit lexique et ne peut donc pas être utilisé dans un autre contexte que celui initialement prévu (p.ex. : un lexique de noms de villes françaises ne peut pas s'appliquer au Canada) on note également que la plupart des études proposent des analyses portant sur différents types de crimes à la fois (p.ex. : Carnaz, Nogueira, Antunes et Ferreira, 2019; Gianola, 2020; Schraagen et al., 2017). Bien que ce type de contribution puisse offrir une portée plus élargie en termes d'applicabilité, la diversité des corpus associés fait en sorte que les algorithmes développés ne peuvent offrir de performances aussi optimales que ceux ayant été développés pour un type de criminalité plus spécifique. À cet égard, quelques études ont porté une attention plus particulière à la détection d'entités nommées dans des documents portant sur des infractions relatives aux stupéfiants (p.ex.: Carnaz, Nogueira, Antunes et Fonseca Ferreira, 2019; Chau et al., 2002), au terrorisme (p.ex.: Inyaem et al., 2009), au crime organisé (p.ex.: Chen, Schroeder, et al., 2003; Ejem, 2017), à la violence contre les femmes (p.ex.: Das et Das, 2017a, 2017b) et aux vols par effraction (p.ex.: Arulanandam et Savarimuthu, 2014; Munasinghe et al., 2014). Ainsi, la création d'une ontologie et de termes spécifiques à une problématique criminelle s'avère appropriée comme en témoignent Petasis et ses collègues (2001) :

« The use of general-purpose resources is ineffective, since in most applications a specialized vocabulary is used, which is not supported by general-purpose lexicons and grammars. For this reason, significant effort is currently put into the construction of generic tools that can quickly adapt to a particular thematic domain. The adaptation of these tools mainly involves the adaptation of domain-specific semantic lexical resources. » (p.426)

De cette façon, il est généralement possible de retrouver une plus grande uniformité dans la structure du texte, dans les documents observés, ainsi que dans le vocabulaire utilisé. Cette précision du corpus sur lequel l'apprentissage est effectué permet habituellement d'obtenir de meilleures performances pour la reconnaissance automatisée (Carnaz, Quaresma, et al., 2019; Ejem, 2017; Schraagen et al., 2017).

La mise en route d'une démarche structurée et d'un modèle algorithmique adaptés pour la REN

Les modèles d'analyse en REN consistent en la mise en place d'un système supervisé d'automatisation pour la reconnaissance d'entités nommées dans un corpus de documents textuels. Ces entités nommées peuvent être hautement diversifiées en fonction des objectifs des analyses qui sont conduites. Les entités les plus communes sont les noms de personnes, d'organisation et de lieux. Ces entités peuvent aussi être précisées, par exemple en ne recherchant que des noms de villes à titre de lieu. À ces entités plus communes, il est également possible d'intégrer à l'algorithme de reconnaissance d'autres types d'entités nommées, comme des dates, des heures, des montants d'argent et des pourcentages (voir Figure 1). Ces différentes reconnaissances peuvent se faire soit par le biais de processus d'apprentissage machine, soit par une reconnaissance basée sur des règles linguistiques précises (rule-based approach) (Chau et al., 2002).

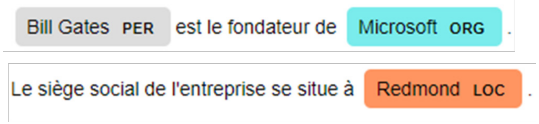

Figure 1 : Visualisation d'une analyse par REN

L'approche par apprentissage machine consiste à « entrainer » un algorithme en lui montrant des exemples du type d'entités que l'on souhaite lui faire reconnaître. Elle est particulière utile lorsque les entités que l'on souhaite reconnaître ne se présentent pas toujours sous la même forme, comme un lieu par exemple, pouvant être une rue, une ville, un pays, ou encore une adresse civique précise. Par le biais de ce procédé, l'algorithme de reconnaissance se construit en fonction d'un certain nombre de caractéristiques sur les entités nommées identifiées, comme les normes morpholexicales, les préfixes et suffixes, ainsi que

la forme de l'entité désignée (ponctuation, nombre, lettre, etc.) (spaCy, 2022). De cette façon, la démarche vise à reconnaître toute entité disposant de caractéristiques similaires à celles ayant été utilisées lors de l'apprentissage, ce qui lui offre beaucoup de flexibilité quant à la nature des entités reconnues. Il convient toutefois de rappeler qu'une plus grande uniformité, notamment entre le corpus d'apprentissage et de reconnaissance, permet une reconnaissance plus exacte.

La reconnaissance basée sur des règles consiste à déterminer les paramètres que l'on cherche à reconnaître en s'appuyant généralement sur des constructions lexicales précises ou des expressions régulières (regex) (Gianola, 2021). Une adresse courriel, par exemple, pourra alors être reconnue simplement par une série de deux chaînes de caractères alphanumériques séparés par une arobase (@), et entrecoupés de points (voir Figure 2).

```
^([a-zA-Z0-9._-]+@[a-zA-Z0-9._-]+\.[a-zA-Z0-9_-]+)$
```

Figure 2 : Expression régulière pour la reconnaissance d'une adresse courriel

Les usages de la REN peuvent donc être très diversifiés. Elles peuvent permettre de réduire un champ de recherche pour en isoler uniquement les entités d'intérêt dans une quantité massive de documents. Une fois ce processus complété, il est possible de classer les documents selon les entités qui y sont nommées, d'en comptabiliser leur fréquence, ou encore de classer les documents en considérant la co-occurrence d'entités de différents types.

Afin d'obtenir des résultats probants pour la REN, il est essentiel de suivre un certain nombre d'étapes préliminaires pour la mise en condition des données textuelles. Ces étapes, présentées à la Figure 3, s'amorcent par l'importation des données, suivi de leur uniformisation et préparation avant la conduite de la REN. Afin de déceler les liens pertinents entre les dossiers et les documents analysés, les entités communes doivent être identifiées et interprétées en prenant en considération le contexte de l'enquête et les limites du protocole mis en œuvre. Dans les paragraphes qui suivent, ces différentes étapes seront abordées plus en détail.

L'acquisition des données

D'abord, lorsque les sources de données diffèrent, il est généralement requis de les convertir sous un même format (p.ex. : convertir les fichiers Word et pdf en texte brut). Les textes ont donc le même format incluant l'encodage approprié. Généralement, les dossiers d'enquête comportent des fichiers de différents formats. On y retrouve des rapports d'événement, des déclarations de témoins, des avis de comparution, des relevés financiers, des registres d'appels téléphoniques, etc. Selon le cas, il peut s'agir de documents en format texte brut, traitement de texte (i.e. : Microsoft Word), pdf, un chiffrier (i.e. : Microsoft Excel), etc. Pour en faciliter l'analyse et le traitement du langage naturel, il est nécessaire de convertir tous ces documents sous un format de texte brut. Plusieurs librairies en langage Python sont disponibles à cette fin et varient en fonction des types de fichiers devant être convertis. C'est le cas entre autres des librairies pdfminer, docx2txt, html2text, etc.

L'uniformisation

Ensuite, afin d'être bien assimilé, le texte brut doit aussi être nettoyé et standardisé. Les documents dans les dossiers d'enquête policière comportent souvent des sections qui ne sont pas pertinentes pour les analyses de REN, comme des en-têtes et pieds de page, des champs administratifs, des numéros de pages, etc. Également, la présence de certains caractères peut nuire aux performances des analyses qui sont conduites. C'est le cas notamment de certaines ponctuations, caractères spéciaux, espaces superflus, etc. Il est donc important de les retirer au préalable, tout en gardant ceux qui pourraient toutefois être nécessaires aux analyses, par exemple les symboles de dollars ($) dans les enquêtes pour fraude. L'ampleur du nettoyage à faire dépend donc des caractéristiques et du contenu des documents à analyser.

Le prétraitement

Puis, un processus que l'on peut généralement diviser en trois grandes étapes d'amorce du traitement du langage naturel est nécessaire pour la réalisation de la REN. Premièrement, le corpus de texte est divisé en phrase, étape que l'on appelle le sentences splitting, sur la base de la ponctuation ou changements de lignes, par exemple. Cette étape servira pour la conduite des étapes ultérieures. Deuxièmement, ces phrases seront encore découpées, mais cette fois en jetons, ou tokens, formés de mots ou de séquences de caractères (Alfred et al., 2014; Asharef et al., 2012; Carnaz, Quaresma, et al., 2019). À cette étape, il s'agit d'isoler les mots, nombres, symboles et espaces, ce qui permettra de faire la liste de chaque composante du texte et d'en calculer entre autres sa fréquence. Une fois l'étape de tokenization réalisée, il faut ensuite attribuer une étiquette grammaticale à chaque jeton en fonction de ses caractéristiques syntaxiques et morphologiques dans la phrase, une étape essentielle pour alimenter l'algorithme de REN. En effet, cette attribution d'étiquettes, aussi appelée part-of-speech tagging, consiste à déterminer si un jeton est un verbe, un nom, un adjectif ou un adverbe, et à identifier ses caractéristiques comme le genre, le nombre et le temps (de verbe surtout). Par exemple, le fait de savoir qu'un jeton constitue un nom dans la

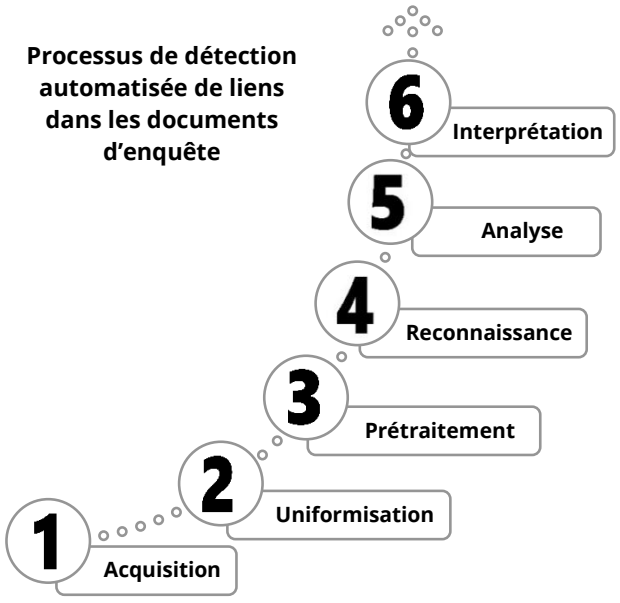

Figure 3 : Démarche structurée pour l'identification de liens par REN

phrase est crucial pour la reconnaissance d'une entité comme un nom de personne.

La reconnaissance d'entités nommées

Aux fins de reconnaissance d'entités nommées à proprement dit, il existe différents algorithmes offrant des performances différentes selon les particularités des documents analysés. Les algorithmes Stanford NER, NLTK, et spaCy figurent parmi les librairies de traitement du langage naturel les plus utilisées et offre des librairies en langage Python. Les librairies de détection d'entités nommées sont implémentées dans différents langages de programmation et utilisent des algorithmes de détections qui sont propres à chacun. SpaCy utilise le langage Python et un algorithme de réseau neuronal afin de détecter les entités nommées (Jafari et al., 2020). NLTK est aussi implémenté en langage Python, mais utilise des classificateurs Naive Bayes à des fins de détections (Xue et al., 2011). Quant au Stanford NLP, il est en langage Java et utilise des champs aléatoires conditionnels (Manning et al., 2014). Les méthodes statistiques utilisées par ces librairies affectent la performance des modèles de détections. C'est ainsi que certains seront en mesure de mieux détecter un type d'entités ou offriront de meilleures détections sur un corpus spécifique. Par exemple, le modèle de base en langue anglaise de Stanford NLP est le meilleur pour détecter les entités de personnes (Schmitt et al., 2019).

En effet, certains algorithmes sont mieux entraînés pour la reconnaissance d'entités spécifiques, ou pour des contextes particuliers, tandis que d'autres sont plus adaptés pour le traitement de documents produits dans une ou des langues particulières. Il est pertinent de mentionner que la plupart des algorithmes sont conçus pour le traitement de texte en langue anglaise. Pour l'analyse de texte en français, sans implication de processus d'apprentissage additionnel, la librairie Polyglot offre un modèle entraîné pour le traitement multilingue, incluant le français. Néanmoins, il est important de considérer que l'usage de tels algorithmes génériques n'offre habituellement pas les meilleurs taux de succès en raison de certaines particularités de la langue.

Pour un usage spécifique en contexte francophone, la reconnaissance d'entités nommées doit souvent se faire dans des phrases longues et complexes, ce qui fait en sorte que le contexte dans lequel les entités peuvent être repérées offre une plus grande diversité (Abeillé et al., 2003). Ainsi, un plus grand nombre de phases d'apprentissage peuvent être nécessaires pour atteindre un même degré de performance que dans la langue anglaise. Ensuite, on retrouve également plusieurs homonymes dans la langue française, ce qui crée des ambiguïtés lors de l'annotation des entités. Par exemple, le mot « neuf » peut à la fois servir d'adjectif pour un nom commun et il peut aussi servir à identifier un chiffre dans le cadre d'une date ou d'une adresse. Dans un autre ordre d'idée, la langue française est aussi composée de plusieurs mots composés, soit des mots qui n'ont pas le même sens, ou n'ont simplement pas de sens, sans la présence d'un autre mot, par exemple « à l'insu » (Abeillé et al., 2003; Petasis et al., 2001).

Afin de pallier ce genre de difficultés, des applications sont spécifiquement conçues pour l'optimisation des algorithmes de REN par apprentissage machine. Typiquement, ce genre d'outil sert à faciliter l'annotation de texte selon les différents types d'entités dont on souhaite faire la reconnaissance, pour ensuite permettre à l'algorithme d'apprendre en fonction des annotations manuelles qui ont été effectuées. Cette démarche nécessite d'abord une reconnaissance d'entités nommées dans un corpus donné où ces dernières ont été identifiées par l'analyste au préalable, de même que la catégorie d'entité dont elles font partie (p. ex. : personne, lieu, organisation, etc.). La quantité de données nécessaire à l'obtention d'un apprentissage machine optimal varie en fonction de la qualité et de la complexité des documents analysés.

Au fil du processus d'apprentissage, il est donc important de mesurer les performances de l'outil de reconnaissance d'entité. Pour ce faire, l'algorithme est testé sur un corpus de validation n'ayant pas été utilisé pour l'apprentissage. Cette validation s'effectue par le calcul d'un F-score s'appuyant sur des indicateurs de précision et de rappel, tel que le montre la Figure 4.

$$\mathbf{F = 2(Précision \cdot Rappel)/(Précision \cdot Rappel)}$$

Figure 4 : Formule de calcul du F-score

L'indicateur de précision suggère à quel point les identifications positives de l'algorithme sont précises, soit en calculant la proportion d'identifications positives effectivement correctes (vrais positifs / (vrais positifs + faux positifs)). Quant à l'indicateur de rappel, il vise à mesurer la proportion de résultats positifs ayant été identifiée correctement par l'algorithme (vrais positifs / (vrais positifs + faux négatifs)). Afin d'améliorer les performances de l'algorithme, d'autres phases d'apprentissages peuvent ensuite être effectuées.

L'analyse des entités communes

Au terme de la démarche de préparation proposée, la REN peut ensuite être conduite sur le corpus de données. Une fois les entités reconnues dans le texte, il est possible d'identifier les entités qui sont communes à plus d'un document ou d'un dossier, selon le besoin de l'analyste. Ces entités communes pourront alors faire office de lien, par exemple dans l'éventualité où une même personne est nommée dans plus d'un dossier d'enquête (Grishman, 2015). Il est toutefois possible, voire fréquent, que les résultats obtenus par l'analyse des entités communes présentent des faux négatifs. En effet, les systèmes de reconnaissance par apprentissage machine ne sont pas infaillibles et il est toujours possible que des entités pouvant constituer un lien entre différents dossiers d'enquête n'aient pas été reconnues. Ce n'est donc pas parce qu'il n'y a pas d'entités communes que les dossiers ne sont pas liés entre eux.

Ce type d'erreur peut aussi survenir lorsque des entités devant être considérées comme les mêmes ne sont pas identifiées de la même façon. Pour que cette mise en relation puisse se faire adéquatement, il est nécessaire de pouvoir fusionner des entités ne présentant pas les mêmes caractéristiques linguistiques. Par exemple, si dans un dossier le suspect principal est identifié comme « John Smith » et que dans un autre ce même suspect est identifié comme « M. Smith », l'unique analyse d'entités communes sans traitement subséquent sur les entités ne permettra pas de déceler le lien entre les deux dossiers. Il en est de même pour les défauts d'orthographe et les fautes de frappe dans les entités reconnues (p. ex. Mark, Makr, Marc, etc.).

Afin de pallier ce problème, une étape subséquente d'analyse de la similitude des entités peut être menée sur l'ensemble des entités reconnues. De cette façon, il est possible de trouver des entités

similaires qui mériteraient d'être fusionnées. Selon le contexte, il sera alors nécessaire de déterminer un seuil de similitude pour l'analyse, soit par la présence d'une partie commune dans différentes entités (p. ex. : John Smith et M. Smith), soit par la présence d'un certain nombre de caractères communs (p. ex. : John, Jhon et Jon), etc. Cette manière de procéder offre la possibilité d'éviter des faux négatifs, mais elle ouvre toutefois la porte à la création de faux positifs, surtout pour certains types d'entités. C'est notamment ce qu'ont constaté Bollé et Casey (2018), qui suggèrent qu'une différence d'un seul chiffre entre deux adresses IP n'offre pas une similitude comparable à une différence d'un seul caractère entre deux noms de personnes : « *For example, a one-bit difference between two IP addresses could actually correspond to different regions. The IP addresses 73.15.110.251 and 73.16.110.251 are both assigned to Comcast Cable, but the first is allocated in California and the second is allocated in Massachusetts.* » (p. S5).

L'interprétation des liens

Une fois les entités communes mises en relation, une interprétation de la nature des entités communes comme des liens pertinents entre les dossiers doit être envisagée et certaines limites doivent également être prises en compte. D'abord, plusieurs entités peuvent être communes entre différents dossiers sans que ceux-ci soient nécessairement liés aux fins d'enquête. La présence d'une adresse commune, par exemple, ne signifie pas forcément que les infractions commises à cet endroit à deux moments distincts sont le fruit des mêmes individus. Il peut notamment s'agir de lieux à forte fréquentation (restaurants, débits de boisson, commerces, etc.) où diverses infractions peuvent avoir été commises. Dans le même ordre d'idée, les entités reconnues peuvent être polysémiques, dans la mesure où elles représentent en réalité des entités différentes. Par exemple, il est fréquent de voir que deux personnes peuvent avoir le même nom, ce qui vient réitérer l'importance d'être prudent lors de la fusion d'entités. Les entités reconnues peuvent également être le fruit de figures de style et porter à confusion. L'utilisation d'une expression telle que « faire la rue Michel » pourrait avoir été identifiée par l'algorithme de REN comme un lieu, ou encore l'expression « tranquille comme Baptiste » comme une personne. Ces entités reconnues ne devraient pas être considérées comme pertinentes pour la mise en relation d'autres dossiers contenant une adresse sur la rue Michel ou la présence d'un individu du nom de Baptiste.

D'ailleurs, l'exercice de passer d'une entité, dans le mon réel, vers une entité nommée, qui s'inscrit dans une ontologie propre à un contexte, demeure une tâche ardue. Une des solutions fréquemment proposées consiste à utiliser un cadre de référence contextuel. Par exemple, dans un corpus particulier, le « Joker » pourrait évoquer la carte à jouer, le personnage ou même le sobriquet d'un des sujets mentionnés dans le rapport d'enquête sous analyse. Ainsi, on conçoit la désambiguïsation comme une tâche intermédiaire (Wilks et Stevenson, 1996), qui permet d'accomplir une série d'autres tâches en langage naturel et qui implique « l'association d'un mot donné dans un texte avec une définition ou une signification (un sens) en le distinguant des autres significations potentiellement attribuables à ce mot » (Ide et Véronis, 1998, p. 3). Or, des chercheurs ont largement utilisé des bases de connaissances comme de Wikipédia comme point de comparaison pour résoudre ce problème (voir entre autres Hahm et al. (2014)). Cette solution est difficilement applicable pour les personnes dans un contexte d'enquête, puisque les personnes ne sont pas connues au sens large. Elle pourrait toutefois compléter ou appuyer la démarche de l'approche par intelligence artificielle, puisque cette dernière n'apprend que le sens probable selon son apprentissage. Ces limites à l'interprétation illustrent les possibilités de faux positifs, renchérissant l'intérêt de traiter minutieusement les résultats obtenus en fonction de leur contexte afin qu'ils apportent une réelle valeur aux enquêtes.

Enfin, il s'avère également pertinent de souligner que le nombre d'entités communes entre deux dossiers distincts ne constitue pas un indicateur de la possibilité que ces entités communes représentent un lien pertinent entre les dossiers. Elles représentent un lien, certes, mais qui ne s'avère peut-être pas d'intérêt dans le contexte de l'enquête. En fait, deux dossiers peuvent contenir des centaines d'entités communes, mais qu'aucune ne constitue un lien significatif dans le contexte de l'enquête, alors que deux autres ne peuvent présenter que très peu d'entités communes, mais témoignant de liens avérés entre ceux-ci. Il est aussi à considérer que les différents types d'entités ne possèdent pas tous le même potentiel indiciaire. Une adresse civique commune entre deux dossiers, par exemple, à davantage de chance d'être non pertinente qu'un numéro de carte de crédit commun dans deux dossiers. Cela dit, cette relation peut varier fortement en fonction du contexte de l'enquête.

Conclusion

L'optimisation des techniques et le recours à l'intelligence artificielle sont aujourd'hui des incontournables au traitement des données massives en contexte d'enquête et de renseignement. Le présent article visait à proposer une démarche en six étapes pour la mise en place d'un protocole structuré d'intégration de la reconnaissance d'entités nommées pour l'identification de liens entre des dossiers d'infraction pour fraude. Bien qu'elle s'appuie sur les capacités informatiques pour contribuer à l'analyse des données, et elle ne prétend pas pouvoir se substituer au travail de l'analyste en renseignement. Principalement, les limites inhérentes à l'interprétation des données issues de ce procédé mettent en lumière toute la pertinence du regard attentif et de la connaissance du contexte criminologique de l'analyste qui demeure essentielle à la réalisation de ce type de tâche. En effet, un tel processus visant la découverte de liens pertinents entre différents dossiers d'enquête ne peut être confié uniquement à un outil de reconnaissance automatisée. Une fois les résultats obtenus, l'analyste doit inévitablement prendre soin d'évaluer quelles entités doivent être fusionnées ou non, lesquelles sont bien ou mal orthographiées, ou encore quelles entités communes s'avèrent pertinentes dans le contexte des dossiers étudiés. Lorsque réalisée adéquatement, l'intégration d'un outil de traitement par REN pour l'analyste en renseignement offre des possibilités inédites pouvant contribuer de manière significative à la conduite d'enquêtes modernes impliquant des quantités massives de données textuelles. Afin d'en connaître davantage sur les contributions et les limites de ce type de démarche, une intégration réelle et des expérimentations concrètes en contexte opérationnelles devraient être réalisées. Cela permettrait du même coup d'évaluer l'impact de l'utilisation de ces nouvelles technologies en contexte policier. Le processus présenté dans la présente étude se concentre sur l'identification d'entités dans des documents d'enquêtes, mais pour que ce processus prenne son sens dans la pratique, il est impératif qu'il soit intégré dans une démarche plus large. Celle-ci viendrait

notamment combler des lacunes importantes des outils disponibles actuellement.

Remerciements

Ce projet a bénéficié du soutien du Conseil de recherches en sciences humaines du Canada (CRSH), financement 892-2020-2007.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for French. Dans A. Abeillé (dir.), *Treebanks* (p. 165-187). Springer. https://doi.org/10.1007/978-94-010-0201-1_10

Alfred, R., Leong, L. C., On, C. K. et Anthony, P. (2014). Malay named entity recognition based on rule-based approach. *International Journal of Machine Learning and Computing*, 4(3), 300-306. https://doi.org/10.7763/IJMLC.2014.V4.428

Arulanandam, R. et Savarimuthu, B. T. R. (2014). Extracting crime information from online newspaper articles. Dans *Second Australasian Web Conference*. https://www.researchgate.net/publication/259932789_Extracting_crime_information_frWom_online_newspaper_articles

Asharef, M., Omar, N. et Albared, M. (2012). Arabic named entity recognition in crime documents. Journal of Theoretical and *Applied Information Technology*, 44(1), 1-6. http://jatit.org/volumes/Vol44No1/fourtyfourth_volume_1_2012.php

Baechler, S., Morelato, M., Roux, C., Margot, P. et Ribaux, O. (2020). Un modèle continu, non linéaire, et collaboratif de l'enquête. *Criminologie*, 53(2), 43-76. https://doi.org/10.7202/1074188ar

Banarescu, A. (2015). Detecting and preventing fraud with data analysis. *Procedia economics and finance*, 32(1), 1827-1836. https://doi.org/10.1016/S2212-5671(15)01485-9

Batura, C. (2021). Applicability of Link Analysis Software in Intelligence Criminal. Dans *Scientia Moralitas Conference Proceedings*. https://www.doi.org/10.5281/zenodo.4762537

Berlusconi, G., Calderoni, F., Parolini, N., Verani, M. et Piccardi, C. (2016). Link prediction in criminal networks: A tool for criminal intelligence analysis. *PLoS ONE*, 11(4), e0154244. https://doi.org/10.1371/journal.pone.0154244

Bollé, T. et Casey, E. (2018). Using computed similarity of distinctive digital traces to evaluate non-obvious links and repetitions in cyber-investigations. *Digital Investigation*, 24(Supplement - Proceedings of the Fifth Annual DFRWS Europe), S2-S9. https://doi.org/10.1016/j.diin.2018.01.002

Brun, O. (2018). Analyste. Dans H. Moutouh (dir.), *Dictionnaire du renseignement* (p. 54-56). Perrin. https://doi.org/10.3917/perri.mouto.2018.01.0054

Bsoul, Q., Salim, J. et Zakaria, L. Q. (2013). An intelligent document clustering approach to detect crime patterns. *Procedia Technology*, 11, 1181-1187. https://doi.org/10.1016/j.protcy.2013.12.311

Carnaz, G., Quaresma, P., Beires Nogueira, V., Antunes, M. et Fonseca Ferreira, N. N. M. (2019). *A Review on Relations Extraction in Police Reports*. Dans Á. Rocha, H. Adeli, L. P. Reis et S. Costanzo (dir.), New Knowledge in Information Systems and Technologies (p. 494-503). Springer. https://doi.org/10.1007/978-3-030-16181-1_47

Chau, M., Xu, J. J. et Chen, H. (2002). Extracting meaningful entities from police narrative reports. *Proceedings of the 2002 annual national conference on digital government research, ICPS Proceedings*(May), 1-5. https://www.diggov.org/library/library/pdf/chau2.pdf

Chen, H., Chung, W., Qin, Y., Chau, M., Xu, J. J., Wang, G., Zheng, R. et Atabakhsh, H. (2003). Crime data mining: An overview and case studies. Dans *National Conference on Digital Government Research*.

Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y. et Chau, M. (2004). Crime data mining: A general framework and some examples. *Computer*, 34(4), 50-56. https://doi.org/10.1109/mc.2004.1297301

Chen, H., Schroeder, J., Hauck, R. V., Ridgeway, L., Atabakhsh, H., Gupta, H., Boarman, C., Rasmussen, K. et Clements, A. W. (2003). COPLINK Connect : information and knowledge management for law enforcement. *Decision Support Systems*, 34(3), 271-285. https://doi.org/10.1016/s0167-9236(02)00121-5

Cofan, S.-M. et Baloi, A.-M. (2017). *Intelligence Analysis: A Key Tool for Modern Police Management - The Romanian Perspective*. Dans J. Eterno, A. Verma, A. Mintie Das et D. K. Das (dir.), Global Issues in Contemporary Policing (p. 165-186). Routledge. https://www.taylorfrancis.com/chapters/edit/10.1201/9781315436975-18/intelligence-analysis-key-tool-modern-police-management%E2%80%94the-romanian-perspective-sorina-maria-cofan-aurel-mihail-b%C4%83loi

Das, P. et Das, A. K. (2017a). Crime Analysis against Women from Online Newspaper Reports and an Approach to apply it in Dynamic Environment. Dans *International Conference on Big Data Analytics and Computational Intelligence*, Chirala, India. https://doi.org/10.1109/icbdaci.2017.8070855

Das, P. et Das, A. K. (2017b). A two-stage approach of named-entity recognition for crime analysis. Dans *8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Delhi, India. https://doi.org/10.1109/icccnt.2017.8203949

De Pauw, E., Ponsaers, P., Bruggeman, W., Van der Vijver, K. et Deelman, P. (2011). *Technology-led policing*. Maklu Publishers. https://biblio.ugent.be/publication/2024677

Deering, T. et Corkill, J. (2017). The intelligence analyst: Attributes, knowledge, skills and characteristics. *Journal of the Australian Institute of Professional Intelligence Officers*, 25(1), 25-39. https://search.informit.org/doi/abs/10.3316/informit.972285516984070

Ejem, R. (2017). *Relation extraction in police records* [Master thesis, Charles University]. http://hdl.handle.net/20.500.11956/90996

Feldman, R. et Dagan, I. (1995). Knowledge discovery in textual databases (KDT). Dans *First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, Montreal, CAN. https://www.aaai.org/Papers/KDD/1995/KDD95-012.pdf

Feldman, R. et Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press. https://doi.org/10.1017/CBO9780511546914

Gianola, L. (2020). *Aspects textuels de la procédure judiciaire exploitée en analyse criminelle et perspectives pour son traitement automatique* [PhD thesis, Université de Cergy-Pontoise]. https://tel.archives-ouvertes.fr/tel-02522680

Gianola, L. (2021). Traitement automatique des langues et linguistique de corpus pour la reconnaissance d'entités en analyse criminelle. *Revue internationale de criminologie et de police technique et scientifique*, 74(3), 363-382. https://www.polymedia.ch/fr/traitement-langues-linguistique-corpus-reconnaissance-entites-analyse-criminelle/

Grishman, R. (2015). *Information extraction*. Dans R. Mitkov (dir.), The Oxford Handbook of Computational Linguistics (2 ed.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199276349.013.0030

Grossrieder, L., Albertetti, F., Stoffel, K. et Ribaux, O. (2013). Des données aux connaissances, un chemin difficile : réflexion sur la place du data mining en analyse criminelle. *Revue internationale de criminologie et de police technique et scientifique*, 66(1), 99-116. https://www.polymedia.ch/fr/des-donnees-aux-connaissances-un-chemin-difficile-reflexion-sur-la-place-du-data-mining-en-analyse-criminelle/

Hahm, Y., Park, J., Lim, K., Hwang, D. et Choi, K.-S. (2014). Named entity corpus construction using wikipedia and dbpedia ontology. Dans *Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. http://www.lrec-conf.org/proceedings/lrec2014/pdf/688_Paper.pdf

Han, J., Kamber, M. et Pei, J. (2012). *Data mining: concepts and techniques*. Elsevier. https://doi.org/10.1016/C2009-0-61819-5

Harper, W. R. et Harris, D. H. (1975). The application of link analysis to police intelligence. *Human Factors*, 17(2), 157-164. https://doi.org/10.1177/001872087501700206

Hassani, H., Huang, X., Silva, E. S. et Ghodsi, M. (2016). A review of data mining applications in crime. *Statistical Analysis and Data Mining*, 9(3), 139-154. https://doi.org/10.1002/sam.11312

Hauck, R. V., Atabakhsh, H., Ongvasith, P., Gupta, H. et Chen, H. (2002). Using Coplink to analyze criminal-justice data. *IEEE Computer*, 35(3), 30-37. https://doi.org/10.1109/2.989927

Hipgrave, S. (2013). Smarter fraud investigations with big data analytics. *Network Security*, 2013(12), 7-9. https://doi.org/10.1016/S1353-4858(13)70135-1

Ide, N. et Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: *The state of the art. Computational Linguistics*, 24(1), 1-40. https://doi.org/https://aclanthology.org/J98-1001.pdf

Inyaem, U., Meesad, P. et Haruechaiyasak, C. (2009). Named-Entity Techniques for Terrorism Event Extraction and Classification. Dans *Eighth International Symposium on Natural Language Processing*, Bangkok, Thaïland. https://ieeexplore.ieee.org/document/5340924

Jafari, O., Nagarkar, P., Thatte, B. et Ingram, C. (2020). SatelliteNER: An Effective Named Entity Recognition Model for the Satellite Domain. Dans *12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management* (IC3K 2020). https://doi.org/10.5220/0010147401000107

Keay, S. et Kirby, S. (2018). The evolution of the police analyst and the influence of evidence-based policing. *Policing: A Journal of Policy and Practice*, 12(3), 265-276. https://doi.org/10.1093/police/pax065

Ku, C. H., Iriberri, A. et Leroy, G. (2008). Crime information extraction from police and witness narrative reports. Dans *IEEE - International Conference on Technologies for Homeland Security*, Boston. https://doi.org/10.1109/THS.2008.4534448

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. et McClosky, D. (2014). The Standford CoreNLP Natural Language Processing Toolkit. Dans *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://aclanthology.org/P14-5010

McCue, C. (2014). *Data mining and predictive analysis: intelligence gathering and crime analysis*. Butterworth-Heinemann. https://doi.org/10.1016/C2013-0-00434-3

McGuire, M. et Holt, T. (2017). *The Routledge Handbook of Technology, Crime and Justice*. Routledge. https://doi.org/10.4324/9781315743981

Merry, S. (2000). *Crime analysis: Principles for analysing everyday serial crime*. Dans D. V. Canter et L. J. Alison (dir.), Profiling property crimes (p. 307-328). Routledge. https://doi.org/10.4324/9781315189192

Milić-Frayling, N. (2005). *Text processing and information retrieval*. Dans A. Zanasi (dir.), Text Mining and its Applications to Intelligence, CRM and Knowledge Management (p. 1-45). WIT Press. https://doi.org/10.2495/978-1-85312-995-7/01

Munasinghe, M., Udeshini, S., Perera, H. et Weerasinghe, R. (2014). Criminal shortlisting and crime forecasting based on modus operandi. Dans *14th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka. https://ieeexplore.ieee.org/document/7083923

O'Connor, C. D. (2021). Thinking about police data: Analysts' perceptions of data quality in Canadian policing. *The Police Journal*, 95(4), 637-656. https://doi.org/10.1177/0032258X211021461

Oatley, G. et Ewart, B. (2011). Data mining and crime analysis. *Wiley Interdisciplinary Reviews: Data mining and knowledge discovery*, 1(1), 147-153. https://doi.org/10.1002/widm.6

Osborne, D. A. (2001). *Four position papers on the role of the crime analyst in policing*. Unplublished MA Social Policy Dissertation. State University of New York, New York.

Petasis, G., Vichot, F., Wolinski, F., Paliouras, G., Karkaletsis, V. et Spyropoulos, C. D. (2001). Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems. Dans *39th Annual Meeting on Association for Computational Linguistics*, Toulouse, FR. https://aclanthology.org/P01-1055.pdf

Piza, E. L. et Feng, S. Q. (2017). The current and potential role of crime analysts in evaluations of police interventions: Results from a survey of the International Association of Crime Analysts. *Police Quarterly*, 20(4), 339-366. https://doi.org/10.1177/1098611117697056

Plouffe, É. (2021, 22 janvier). *Deux fois plus de victimes de fraude au Canada en 2020*. Radio-Canada. https://ici.radio-canada.ca/nouvelle/1764924/fraude-centre-antifraude-grc-pandemie

Rossy, Q. (2011). *Méthodes de visualisation en analyse criminelle: approche générale de conception des schémas relationnels et développement d'un catalogue de patterns* [Université de Lausanne]. https://serval.unil.ch/resource/serval:BIB_1AC0D89CA5A4.P001/REF.pdf

Rossy, Q. (2016). *La visualisation relationnelle au service de l'enquête criminelle*. Dans R. Boivin et C. Morselli (dir.), Les réseaux criminels (p. 17-50). Presses de l'Université de Montréal. https://www.pum.umontreal.ca/catalogue/les-reseaux-criminels

Rossy, Q., Décary-Hétu, D., Delémont, O. et Mulone, M. (2018). *The Routledge International Handbook of Forensic Intelligence and Criminology*. Routledge. https://doi.org/10.4324/9781315541945

Rossy, Q., Ribaux, O., Boivin, R. et Fortin, F. (2019). *Le traitement de l'information dans l'enquête criminelle*. Dans M. Cusson, O. Ribaux, É. Blais et M. M. Raynaud (dir.), Nouveau traité de sécurité. Sécurité intérieure et sécurité urbaine (p. 428-446). Editions Hurtubise. https://editionshurtubise.com/livre/nouveau-traite-de-securite/

Schmitt, X., Kubler, S., Robert, J., Papadakis, M. et LeTraon, Y. (2019). A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. Dans *Sixth International Conference on Social Networks Analysis, Management and Security* (SNAMS). https://doi.org/10.1109/SNAMS.2019.8931850

Schraagen, M., Brinkhuis, M. et Bex, F. (2017). Evaluation of Named Entity Recognition in Dutch online complaints. *Computational Linguistics in the Netherlands Journal*, 7, 3-16. https://dspace.library.uu.nl/handle/1874/356185

Schroeder, J., Xu, J., Chen, H. et Chau, M. (2007). Automated criminal link analysis based on domain knowledge. *Journal of the American Society for Information Science and Technology*, 58(6), 842-855. https://doi.org/10.1002/asi.20552

spaCy. (2022). API: EntityRecognizer. https://spacy.io/api/entityrecognizer

Tan, A.-H. (1999). Text mining: The state of the art and the challenges. Dans *PAKDD Workshop on Knowledge Discoverery from Advanced Databases*, Beijing, China. https://www.researchgate.net/publication/2471634_Text_Mining_The_state_of_the_art_and_the_challenges

Vossen, P. (2002). WordNet, EuroWordNet and Global WordNet. *Revue française de linguistique appliquée*, VII, 27-38. https://doi.org/10.3917/rfla.071.0027

Westphal, C. (2008). *Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies*. CRC Press. https://doi.org/10.1201/9781420067248

Wilks, Y. et Stevenson, M. (1996). *The Grammar of Sense: Is word-sense tagging much more than part-of-speech tagging?* (publication no CS-96-05). https://doi.org/10.48550/arXiv.cmp-lg/9607028

Xue, N., Bird, S., Klein, E. et Loper, E. (2011). Natural Language Processing with Python. *Natural Language Engineering*, 17(3), 419-424. https://doi.org/10.1017/S1351324910000306