Extraction et caractérisation de noyaux d'événements liés à la pollution industrielle

Chuanming Dong, Philippe Gambette, Catherine Dominguès

Extraction et caractérisation de noyaux d'événements liés à la pollution industrielle

Chuanming Dong[1], Philippe Gambette[2], Catherine Dominguès[3]

[1]LASTIG, Univ Gustave Eiffel, ENSG, IGN et ADEME – chuanming.dong@ign.fr
[2]Univ Gustave Eiffel, CNRS, LIGM – philippe.gambette@univ-eiffel.fr
[3]LASTIG, Univ Gustave Eiffel, ENSG, IGN – catherine.domingues@ign.fr

Abstract
We seek to characterize events related to industrial pollution, which are extracted from administrative texts. Each event can be characterized by a trigger that reflects its nature (administration, depollution, etc.). We propose a typology to categorize these triggers and a hybrid method to automatically perform this categorization. This method combines a machine learning approach (based on word embeddings and supervised learning) with a textometric approach (based on the visualization of the co-occurrence of frequent words). It allows characterizing event triggers automatically with an f-score of 80%.

Keywords: statistical analysis, natural language processing, machine learning, clustering, tree cloud, textometry

Resumè
Nous cherchons à caractériser des événements liés à la pollution industrielle, extraits de textes administratifs. Chaque événement peut être caractérisé par un déclencheur qui rend compte de sa nature (administrative, de dépollution, etc.). Nous proposons une typologie pour catégoriser ces déclencheurs et une méthode hybride pour réaliser automatiquement cette catégorisation. Cette méthode combine une approche d'apprentissage automatique (fondée sur des plongements de mots et un apprentissage supervisé) avec une approche textométrique (fondée sur la visualisation de la cooccurrence des mots fréquents). Elle permet de caractériser les déclencheurs d'événements automatiquement avec une f-mesure de 80%.

Mots clés : analyse statistique, traitement automatique des langues, apprentissage automatique, clustering, nuage arboré, textométrie

1. Introduction

La pollution est une des préoccupations centrales des français, et les activités industrielles constituent une source potentielle et importante de pollution. En

France, c'est le ministère en charge de l'Écologie qui collationne et diffuse les informations concernant les activités industrielles et les sites sur lesquels elles ont été ou sont exercées, par l'intermédiaire des bases de données BASOL, BASIAS et S3IC. En particulier, BASOL inventorie les sites et sols pollués, ou potentiellement pollués, appelant une action des pouvoirs publics, à titre préventif ou curatif.

Dans le cadre normal de leurs activités, les industries manipulent ou produisent des produits éventuellement toxiques. Pour chaque entreprise, BASOL rassemble dans différents champs les informations concernant l'activité industrielle et les produits manipulés. Cependant, des événements peuvent aussi se produire qui ne concernent pas l'activité industrielle normale et peuvent être à l'origine de pollutions dues par exemple à une quantité anormale de certaines substances ou à leur contact accidentel. Ces événements sont exceptionnels par rapport à l'activité industrielle normale, et à ce titre non associés à des champs de la base; ils sont alors décrits de manière textuelle et ces textes figurent dans la base, contribuent à la description de l'entreprise mais, non structurés, ne sont pas interrogeables par l'intermédiaire de l'interface d'interrogation et de visualisation fournie sur le site officiel. Pourtant, ces textes contiennent des informations cruciales pour évaluer la situation des sites industriels, en particulier pour connaître les pollutions déjà constatées ou prévisibles ainsi que les actions de dépollution déjà mises en œuvre ou à prévoir. Ces informations intéressent, à divers titres, à la fois les citoyens et les experts. Nous visons donc à construire une mémoire des sites industriels, sous la forme d'une base de données unique (BDU) décrivant les activités normales et les événements exceptionnels concernant les sites industriels, en appariant et réorganisant les informations structurées contenues dans plusieurs bases de données, (dont BASOL) et en extrayant, des textes descriptifs contenus dans les mêmes bases, les caractéristiques des événements survenus sur les sites industriels.

Les éléments permettant de caractériser un événement industriel (noyau, activité, entreprise, polluant, etc.) ainsi que la méthode d'extraction de ces éléments dans les textes descriptifs ont été décrits dans (Dong *et al*., 2021). L'examen de ces caractéristiques a montré la diversité des événements potentiels. D'autre part, la restitution au public du contenu de la BDU, sous une forme la plus complète possible mais aussi intelligible, nécessite de catégoriser et hiérarchiser les informations. En particulier, tous les événements n'ont pas la même importance quant à l'évaluation de la pollution d'un site. L'objet de cet article est de présenter une typologie des événements industriels et les outils qui permettent de classer dans cette typologie, de la manière la plus automatique possible, les événements industriels identifiés dans les textes descriptifs. La démarche proposée s'appuie sur un corpus textuel de descriptions d'événements industriels issues de l'extraction de (Dong *et al*. 2021) et la caractérisation d'un événement pertinente

pour cette thématique. D'une part, elle propose une typologie des événements permettant de classer leurs noyaux en plusieurs classes sémantiques. D'autre part, elle combine une approche textométrique et des méthodes et outils d'apprentissage automatique pour classer dans cette typologie des mots caractéristiques de chaque classe, extraits de ces noyaux, que nous appelons « déclencheurs », et pour enrichir cet ensemble de déclencheurs. La mise en forme des données (préparation de la base de données BASOL et caractéristiques des événements industriels) est rappelée dans la section suivante. La construction de la typologie de ces événements est expliquée dans la section 3. La section 4 expose la démarche de classification des événements industriels dans cette typologie et d'augmentation du vocabulaire désignant les événements. Une conclusion ainsi que des limites de la démarche sont présentées dans la dernière section.

2. Mise en forme des données

Définie dans les années 1990 et renseignée par les directions régionales de l'Environnement, de l'Aménagement et du Logement, BASOL est une base de données publique qui vise à recueillir les informations sur les sites industriels potentiellement pollués. Chaque enregistrement correspondant à un site contient, outre des données structurées caractérisant l'entreprise et son activité, un texte qui décrit des événements relatifs à l'activité normale, à des incidents et des accidents conduisant éventuellement à des pollutions, à l'aménagement ou au réaménagement du site, etc. En 2020, BASOL contenait les informations de plus de 7 000 sites industriels en France. Les textes descriptifs de BASOL constituent le corpus de travail pour l'identification et l'analyse des événements industriels. Il est caractérisé par une langue soutenue et l'emploi d'un vocabulaire spécifique de ces événements industriels, et contient 155 587 phrases, soit 48 032 mots, dont des exemples suivent :

Un diagnostic initial et une évaluation simplifiée des risques (ESR) ont été prescrites par arrêté du 17 avril 2000 conformément à la circulaire du 3 avril 1996.

Les terres polluées correspondant aux zones P9S21 et PY19B2 ont été excavées et évacuées vers des installations de stockage de déchets et des centres de traitement biologique des terres en 2001 et 2002.

Un événement industriel est défini comme une action qui a eu lieu et a modifié la situation de l'entreprise. Les exemples précédents montrent que l'événement est caractérisé par un noyau (souligné) qui définit le type de l'événement, administratif (avec le déclencheur *prescrit*) dans le premier exemple et de réhabilitation (avec les déclencheurs excavées et *évacuées*) dans le deuxième. On s'intéresse dans cet article à un corpus constitué par les noyaux d'événements. Ce corpus a été construit en annotant automatiquement le corpus BASOL à l'aide

du modèle d'annotation automatique entraîné sur une portion de ce même corpus (Dong *et al*., 2021).

3. Caractérisation et typologie des déclencheurs d'événements

La diversité des déclencheurs ainsi que leurs conséquences différenciées sur la situation de pollution d'un site industriel nécessitent de construire une typologie de ces déclencheurs qui permettra de les regrouper sur des critères sémantiques et facilitera l'interrogation de la BDU. Afin de construire cette typologie, une méthode fondée sur la construction d'un nuage arboré des noyaux d'événements (Gambette et Véronis, 2009) a été mise en place. Elle permet de faire apparaître les mots les plus fréquents au sein de ces noyaux, organisés dans un arbre qui rapproche les mots apparaissant fréquemment proches l'un de l'autre au sein d'un noyau. L'observation de ce nuage arboré[1] permet d'identifier six classes d'événements qui constituent le point de départ de la typologie :

- **admin** : les actions provenant de l'autorité administrative (autorisation, demande, prescription, etc.) ;
- **nuis** : les actions décrivant des nuisances contre l'environnement ;
- **diag** : les actions techniques concernant le diagnostic de pollution ;
- **depol** : les traitements relatifs à la réhabilitation ;
- **indus** : le parcours administratif de l'entreprise (implantation, changement ou cessation d'activité, fermeture, etc.) ;
- **neutre** regroupe les actions qui ne peuvent pas être classées dans les classes précédentes, en particulier celles qui, comme : réaliser, effectuer, mettre en œuvre, opérer etc. annoncent la réalisation d'une autre action qui pourra être classée dans la typologie.

On peut donc créer une liste de mots simples spécifiques à chaque classe pour obtenir un premier ensemble de déclencheurs. Contrairement à une méthode d'apprentissage supervisé qui nécessite de connaître préalablement la liste des étiquettes, l'utilisation du nuage arboré permet d'identifier les étiquettes pertinentes. Cette méthode fondée sur les mots fréquents et leur cooccurrence est particulièrement adaptée dans notre situation où le sens des noyaux d'événements est fortement lié à leur contenu lexical.

4. Enrichissement des déclencheurs des noyaux

La capacité de représentation des nuages arborés est limitée à une centaine de mots les plus fréquents. Ceci ne permet de repérer qu'un nombre limité de déclencheurs, et pénalise les déclencheurs moins fréquents, et comme on peut

[1] voir ce nuage arboré dans le matériel supplémentaire de cet article, incluant le code Python développé et les données : https://github.com/DongChuanming/JADT2022_classification_et_enrichissement

l'observer avec la catégorie nuis, les catégories les moins fréquentes de la typologie (un problème également présent dans Gambette *et al*., 2018). Nous combinons donc deux approches pour augmenter le vocabulaire des déclencheurs d'événements classés dans la typologie. D'une part, une démarche itérative consiste à répéter l'extraction manuelle de mots depuis un nuage arboré représentant les noyaux d'événements ne contenant aucun des mots extraits auparavant. D'autre part, une démarche d'apprentissage automatique, à l'aide d'un algorithme de classification fondé sur des plongements lexicaux de mots extraits de ces nuages arborés, est adoptée pour capter également des termes peu fréquents.

4.1. Protocole d'enrichissement à partir des mots fréquents

Le peuplement de la typologie comporte deux axes : reconnaître les formes fléchies des déclencheurs déjà catégorisés (enrichissement horizontal), et reconnaître les nouveaux termes qui n'ont pas le même lemme que les déclencheurs déjà catégorisés (enrichissement vertical).

Le procédure de peuplement est itérative et chaque itération comporte deux étapes. La première étape, déjà mentionnée dans la section 3, vise à réaliser l'enrichissement vertical et consiste à obtenir un sac de mots (S1) en extrayant les déclencheurs manuellement à partir du nuage arboré dans les catégories de typologie. À la deuxième étape (pour réaliser l'enrichissement horizontal), une liste (L1) est générée pour chaque classe en ajoutant certaines formes dérivées manuellement (nom prédicatif si le verbe est un déclencheur, verbe à l'infinitif si le nom prédicatif est un déclencheur) des déclencheurs de cette classe, ainsi que les formes fléchies de tous ces termes, générées automatiquement à partir du dictionnaire DELA (Paumier, 2011). À l'itération suivante, un nouveau nuage arboré est construit à partir des noyaux d'événements ne contenant aucun déclencheur de la liste S1+L1. De même qu'à l'itération précédente, un nouveau sac de mots (S2) et les listes correspondantes de leurs formes dérivées et fléchies (L2) sont générés. Le processus itératif se termine quand le nombre de déclencheurs ajoutés n'augmente plus significativement.

Cette méthode permet donc non seulement de capter des termes un peu moins fréquents dans une démarche manuelle mais aussi de tester et d'alimenter une procédure d'enrichissement automatique à partir d'un classifieur, tout en priorisant l'enrichissement des classes qui contiennent le moins de termes.

4.2. Protocole d'enrichissement à partir d'un classifieur

Simultanément avec l'enrichissement manuel à partir des nuages arborés, une approche d'apprentissage automatique est mise en œuvre pour automatiser la procédure de classification des déclencheurs et finalement enrichir la liste des déclencheurs. À la fin de la n-ième itération, un classifieur est entraîné sur

l'ensemble (S1+...+Sn)+(L1+...+Ln) des déclencheurs déjà catégorisés, et est évalué sur l'ensemble Sn+1+Ln+1 des déclencheurs catégorisés à l'itération suivante. Les données d'apprentissage sont donc accumulées à chaque itération. En combinant les sacs de mots et les listes des formes fléchies, le classifieur est entraîné et évalué à la fois sur les enrichissements horizontal et vertical. La classification supervisée est fondée sur le perceptron multicouche (Haykin, 1994), et les plongements de mots sont fournis par le modèle de langue française CamemBERT (Martin *et al*., 2020).

Finalement, l'implémentation d'une méthode itérative permet de mieux cibler les efforts : ajuster le corpus d'apprentissage en ajoutant au fur et à mesure les déclencheurs les moins fréquents, raffiner le classifieur en intégrant les plongements de ces nouveaux déclencheurs. Ainsi, le classifieur obtenu à la fin de la première itération obtient, pour la classe nuis, une précision de 0,77 mais seulement 0,30 pour le rappel. En conséquence, dans l'apprentissage de la 2ème itération, le nombre des lemmes déclencheurs de la classe nuis est augmenté par rapport à ceux des autres classes, pour renforcer la flexibilité du classifieur dans la reconnaissance de cette classe. Ainsi le modèle est amélioré, obtenant à la fin de l'itération 2, une précision de 0,97 et une rappel de 0,70 pour cette même classe. Pour l'ensemble des classes, la f-mesure passe de 0,51 à la fin de la première itération à 0,80 à la fin de la deuxième[2].

5. Conclusion

Dans cet article, nous avons présenté une approche pour construire et enrichir une typologie sémantique concernant les actions de déclencheurs dans les événements industriels, qui combine les méthodes textométriques et statistiques. En combinant l'apprentissage automatique avec la classification manuelle itérative, une méthode d'apprentissage supervisé itératif est conçue, et celle-ci permet, en se fondant d'abord sur le vocabulaire le plus fréquent dans chaque classe, de créer rapidement des données d'apprentissage de grande taille, susceptibles d'être enrichies en utilisant un modèle de langue fondé sur des plongements. Un atout du résultat obtenu est que chaque classe thématique d'intérêt est caractérisée par un ensemble de déclencheurs susceptibles d'être présentés à une personne non experte du domaine pour l'aider à interroger et explorer la BDU. En revanche, cette approche n'a pas pour objectif d'optimiser à la fois la précision et le rappel de la classification des noyaux, qui permettrait des requêtes ciblées sur certaines classes plus précises de noyaux ou de

[2] détails des évaluations et liste de déclencheurs enrichie par la méthode itérative et le classifieur disponibles ici : https://github.com/DongChuanming/JADT2022_classification_et_enrichissement

déclencheurs, avec garantie d'exhaustivité des résultats. Le ciblage des déclencheurs les plus fréquents pour constituer le corpus d'apprentissage permet d'assurer un niveau minimum de représentativité d'apprentissage, les déclencheurs moins fréquents, absents du corpus d'apprentissage, pouvant être repérés par le calcul de plongements. Une limite de la découverte et du classement des déclencheurs provient de la segmentation en tokens (mots ou portions de mots) pour le calcul des plongements, peu propice à l'étiquetage d'expressions polylexicales.

Ultérieurement, le classifieur sera utilisé pour classer des déclencheurs déjà extraits du corpus BASOL. Le principal problème à régler est le cas où un noyau d'événement contient plusieurs déclencheurs de catégories différentes, comme dans les noyaux stopper l'extension de la pollution (catégories dépol et nuis) et avait prescrit un audit sécurité (catégories admin et dépol). Plusieurs pistes sont envisagées, par exemple établir un ordre de priorité entre les catégories. Finalement, les déclencheurs d'événement, caractérisés par la typologie, pourront être enregistrés dans la BDU avec leurs composants extraits du corpus BASOL.

Remerciements

Ce travail est en partie financé par l'Agence de l'Environnement et de la Maîtrise de l'Energie (ADEME), 20, avenue du Grésillé- BP 90406 49004 Angers Cedex 01 France.

Bibliographie

Dong C., Gambette P. et Dominguès C. (2021). Extracting Event-related Information from a Corpus Regarding Soil Industrial Pollution. In Proc. of KDIR 2021, vol. 1, pp. 217-224.

Gambette P., Kyriacopoulou T., Lechevrel N. et Martineau C. (2018). *Anatomie, animaux, vocabulaire de la vivisection. Construire des ressources lexicales pour visualiser une thématique dans un corpus littéraire*. In Animalhumanité. Expérimentation et fiction : l'animalité au cœur du vivant, p. 223-232.

Gambette P. et Véronis, J. (2010). Visualising a Text with a Tree Cloud. In Proc. of IFCS 2009, Studies in *Classification, Data Analysis, and Knowledge Organization* 40, pp. 561-570.

Haykin, S., & Lippmann, R. (1994). Neural networks, a comprehensive foundation. *International journal of neural systems*, 5(4), 363-364.

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., de la Clergerie É., Seddah D. et Sagot B. (2020). CamemBERT: a tasty French language model. In Proceedings of the *58th Annual Meeting of the Association for Computational Linguistics*, pp. 7203-7219.

Paumier S. (2011). *Unitex - Manuel d'utilisation*, https://hal.archives-ouvertes.fr/hal-00639621/document