Détection d'événements géo-chrono-localisés sur Twitter

Hosni Seffih [1, 2], **Myriam Lamolle** [2], **Aurélie Pradelles** [1], **Zhen Wang** [1], **Jérémy Lhez** [1]

1. *GEOLSemantics, 12 avenue Raspail, 94250 Gentilly*
   prenom.nom@geolsemantics.com
2. *LIASD - IUT de Montreuil - Université Paris 8, 140 Rue de la Nouvelle France, 93100 Montreuil*
   m.lamolle@iut.univ-paris8.fr

RÉSUMÉ.
Dans cet article, nous présentons un processus complet d'extraction d'événements géo-chrono-localisés dans des tweets sous forme de triplets (temps, lieu, action), mais aussi une technique de regroupement de tweets parlant d'un même événement afin d'apporter au fur et à mesure du processus de détection plus de précision sur la date et le lieu de l'événement. À cette fin, notre système applique une analyse linguistique généraliste fondée sur des règles linguistiques ainsi qu'une extraction sémantique guidée par une ontologie d'événements. Ces deux phases ont été adaptées pour l'extraction d'information dans le contexte de Twitter où les messages sont courts, les abréviations et fautes d'orthographe nombreuses. Un échantillon de tweets collectés récemment sera testé sur notre système et sur un système d'apprentissage statistique (i.e. Machine Learning).

ABSTRACT.
In this paper, we present the complete process of extracting geo-chrono-localized events from tweets. We have designed a system capable of detecting an event by a triplet (time, place, action), and the first published tweet concerning it if possible. We also show how tweets are grouped talking about the same event in order to bring more precision during the detection process of event date and location. Our system applies general linguistic analysis based on linguistic rules, and a semantic extraction guided by an events ontology. The two phases have been adapted for information extraction in the context of Twitter where messages are short and contain many abbreviations and misspellings. In order to demonstrate the effectiveness of our approach, a sample of tweets will be tested on our system and a machine learning system.

MOTS-CLÉS : *Twitter, Système d'information, Détection d'événements géo-chrono-localisés, Regroupement de tweets, Ontologie, Machine learning*

KEYWORDS: *Twitter, Information system, Detection of geo-chrono-localized events, tweets' merging, Ontology, Machine learning*

1. Introduction

De façon simplifiée, un système d'information (SI) peut être vu comme un ensemble de sources de données et d'applications permettant la gestion d'information. Avec l'avènement du Web et de l'*open data*. Twitter représente une des sources externes, très utilisées, mais nécessitant une adaptation des ressources propres au SI traitant habituellement des textes de grande taille pour rechercher des informations particulières ; d'autant plus lorsque une fonctionnalité du SI est la détection d'événements à des fins d'alerte de danger (par exemple, un feu, un tremblement de terre, etc.).

La détection d'événements « géo-chrono-localisés » n'est pas chose facile sur Twitter sachant qu'elle doit se faire au plus tôt donc en un minimum de tweets (Ying *et al*., 2018). De façon générale, beaucoup d'ambiguïtés peuvent survenir des termes utilisés s'ils ne sont pas replacés dans leur contexte. Par exemple, le terme « Nice » peut faire référence à la ville du sud de la France mais aussi à la notion de « joli » (selon l'anglais) alors que la conversation est en français. Autant nous sommes capables intuitivement de faire cette distinction et donc de localiser ou non l'événement, autant cela devient difficile pour une machine ou un programme. Dans le cas des réseaux sociaux, et plus précisément dans les tweets, le vocabulaire et la tournure des phrases sont encore plus succincts puisque contraints à 144 caractères jusqu'à récemment, en 2017 passage à 280 caractères (Perez, 2017). Par exemple, il apparaît beaucoup plus de mots contractés ("tkt" pour "Ne t'inquiète pas"), de phonétisations ("foto" pour "photo" ou "C" pour "c'est"), de pictogrammes, etc. Quant à la détection d'événements, nous pouvons être guidés par les *hashtags* mais cela ne nous permettra pas de déterminer où et quand se déroule(-era) l'événement. Ces informations sont essentielles lors du déclenchement d'alertes automatiques pour des risques climatiques, des mises en sécurité de personnes, etc., en temps réel ou semi-réel.

Cet article s'inscrit dans cette perspective pour, d'une part, extraire et améliorer le plus possible la sémantique d'un tweet, et, d'autre part, augmenter la pertinence des alertes par une datation et une précision de localisation acceptables. Concernant notre premier objectif, les outils d'extraction sont fondés sur une analyse morphosyntaxique profonde généraliste puis sur une extraction sémantique détectant les concepts décrits dans l'ontologie de l'application grâce à des règles d'extraction utilisant les résultats de l'analyse morphosyntaxique profonde. Afin d'affiner la pertinence du déclenchement de l'alerte, notre deuxième objectif, nous allons nous attacher à compléter une ontologie existante dont la modélisation va être dirigée par un corpus *ad hoc* au contexte des événements afin d'inférer des motifs (dits *patterns*) de « cause à effet », problème assez complexe selon (Besnard *et al*., 2008) qui ont présenté un ensemble de modèles d'inférence formels, depuis des déclarations causales jusqu'aux déclarations explicatives. Ces modèles présentent des prémisses ontologiques qui sont considérées comme essentiels pour déduire les déclarations explicatives. Le domaine d'application étant l'extraction d'événements demandant une intervention des autorités locales, nous limitons notre étude à des événements présents ou du passé très récent.

Dans le cadre du projet "Safecity", GEOLSemantics réalise un système capable de détecter dans les réseaux sociaux (en particulier Twitter) des événements (incendie, accidents, inondation, etc.) qui pourraient nécessiter l'intervention des autorités locales (pompier, ambulance, police, etc.), et de caractériser ces événements en particulier avec leur localisation précise. À la demande des partenaire, nous devons envoyer l'alerte via Kafka [1] dans un contexte de *Stream processing*.

L'article est structuré comme suit : après un état de l'art sur le traitement des tweets notamment pour la détection des événements et des lieux, nous allons présenter notre processus de détection d'événements. La section suivante décrit les différentes briques logicielles sur lesquelles repose notre processus. Il s'ensuit une section concernant l'expérimentation proprement dite dans laquelle nous comparerons la partie détection d'événement de notre système avec un système d'apprentissage statistique. Nous finirons par une conclusion générale et les perspectives envisagées.

2. État des lieux sur la détection d'événements

Twitte est un site de microblogging, qui a permis, entre autres, de démocratiser la diffusion de l'information (Khamis *et al*., 2017), même si cela comporte un risque de propagation de fausses nouvelles (Guibon *et al*., 2019). L'orthographe liée à l'écriture des messages s'est bien améliorée depuis les premiers tweets, mais il reste que ces messages sont souvent écrits phonétiquement, avec des mots collés et/ou des abréviations (140 signes passés à 280 depuis novembre 2017). Les outils d'analyse linguistique et d'extraction sémantique doivent être adaptés pour être robustes à ce type de rédaction. Cela se traduit par des interprétations de mots plus ambiguës et par une syntaxe plus simple avec un manque de mots-outils mais, cependant, dans un ordre plus standard et avec une construction simplifiée des phrases.

Les deux problèmes majeurs rencontrés lors de l'extraction d'événements à partir des tweets sont l'imprécision temporelle (la date/heure de l'événement n'est pas mentionnée systématiquement, ou des expressions de date/heure relatives sont utilisées comme « vient de », « va organiser », « demain », etc.) et l'imprécision géographique (« pas loin de », « devant », etc.).

Concernant la détection des dates dans les tweets, par exemple, pour la prédiction d'événements, en se basant sur une approche d'analyse sémantique automatique et sur des traitements NLP [2] sur des tweets, et en utilisant une modélisation linéaire, (Wang *et al*., 2012) ont pu prédire des délits de fuite. (Aramaki *et al*., 2011) ont pu démontrer que des instruments NLP peuvent permettre l'extraction d'informations pertinentes, en utilisant un classifieur SVM [3] sur un corpus de tweets sur la grippe. (Sakaki *et al*., 2010) ont pu détecter en temps réel des tremblements de terre et lancer une alerte en

1. https://kafka.apache.org/
2. Natural Language Processing
3. Support Vector Machine

avertissant des utilisateurs enregistrés, alerte émise bien avant celle de la JMA (Japan Meteorological Agency).

La détection des lieux, quant à elle, se fait sur quelques paramètres à savoir le tweet de l'utilisateur, son réseau (*followers*, *following*) et ses métadonnées (*profile location*, *tweet timezone*). Selon (Hecht *et al*., 2011), 34% des lieux déclarés sont des lieux sarcastiques. Twitter permet aux utilisateurs d'afficher leurs coordonnées GPS mais seulement 1% des utilisateurs le font vraiment selon (Han *et al*., 2013). Nous avons vérifié cet état de fait sur environ 1000 tweets extraits sur la région parisienne par *bounding box*, c'est-à-dire que les utilisateurs permettent à l'application de se connecter à leur GPS. Ce test a révélé que 19% des tweets n'avaient pas déclaré de lieu dans *profile location*, 38% ont déclaré un autre lieu, 16% ont déclaré un lieu inexistant (cas des *troll*) et seulement 28% ont déclaré la bonne localisation, des enseignes de commerce pour la plupart. D'autre part, nous avons aussi réalisé un autre test en extrayant 1000 tweets commençant par *"Je suis à Paris"*. Dans ce cas, 38% des utilisateurs n'avaient pas déclaré le lieu, 35% avaient déclaré un autre lieu, 20% étaient des trolls et seulement 8% avaient déclaré la bonne géolocalisation Paris. Le lieu déclaré dans le tweet est considéré comme une valeur sûre pour la détection du lieu par rapport à la géolocalisation de Twitter. Dernièrement, (Huang, Carley, 2019) a proposé une géolocalisation hiérarchique pour des utilisateurs de Twitter, fondé sur un réseau de neurones.

Ces deux détections sont essentielles pour détecter des événements, notamment lors de la gestion des catastrophes naturelles (Kryvasheyeu *et al*., 2015) ou le suivi d'une épidémie (Broniatowski *et al*., 2013) entre autres. Pour (Hoang, Mothe, 2018) les événements sont représentés par trois dimensions principales: le temps, le lieu et les informations relatives à l'entité, ils ont fait des travaux sur la relation rappel/précision, combinant différentes méthodes pour extraire les lieux. En ce qui nous concerne, nous visons essentiellement à identifier les événements contenant au minimum une indication de lieu et une temporalité au présent ou au passé proche, exprimés dans le contenu du message, afin de lancer des alertes. Les informations relatives à l'événement permettent de le caractériser par des précisions sur qui (personne ou organisation) fait quoi (action, événement, expérience), quand (date), où (lieu), avec quoi (objet) et combien (quantité ou mesure). Ceci devrait permettre de lever l'ambiguïté des entités et des actions grâce à la définition des contextes sémantiques.

D'autre part, un tweet ne peut pas toujours refléter à lui seul l'importance d'un événement. Il faut être capable d'identifier qu'un autre message relate le même événement ou non (Jackoway *et al*., 2011). C'est indispensable pour ne pas « noyer » l'alerte par des informations redondantes. Il faut aussi être capable de compléter l'information au fur et à mesure que de nouveaux messages arrivent. Enfin, en ce qui concerne des événements annoncés, il faut pouvoir les suivre pour confirmer s'ils arrivent ou non, s'ils auront lieu ou non. En suivant les méthodes orientées caractéristiques (*feature-pivot methods*) qui sont plus sémantiques car elles étudient la distribution des mots et extraient des événements selon un motif de mots défini (dit *pattern*) (Kleinberg, 2003), un événement précis pourrait être défini. Un événement est donc représenté par un

certain nombre de mots-clés avec un nombre d'apparition (Weng, Lee, 2011). Nous pouvons aussi observer que la détection d'un événement dans un lieu et une date précis nous incite à avoir une micro-détection de la date, du lieu et de l'action avant de passer à la macro-détection de l'événement.

3. Processus de détection d'événements

Nous allons maintenant aborder le processus qui permet de détecter chaque élément constitutif d'un événement E pour lancer une alerte, à savoir A = action, L = Lieu, D = Date. Ces trois éléments vont nous permettre de regrouper des tweets parlant du même événement, et ce, même si dans certains tweets, tous les éléments ne sont pas renseignés. Il s'agit donc d'obtenir : $E = (A, L, D)$ où $A = f_{action}(\{tweet\}_n)$, $L = f_{lieu}(\{tweet\}_n)$, $D = f_{date}(\{tweet\}_n)$ et f représente la fonction de raffinement d'un des critères $action$, $lieu$, $date$. Notons que $lieu$ peut être un lieu précis ou une zone géographique, et $date$ une date précise ou un intervalle de temps. La figure 1 présente l'architecture de notre approche constituée de trois parties : la collecte des tweets et l'extraction d'information en deux dimensions (partie de gauche), la chaîne de traitement GEOL en quatre étapes (partie centrale) et la sortie de l'événement formalisé pour l'envoi d'alerte en JSON (partie de droite) dont le détail est donné dans les sous-sections suivantes.

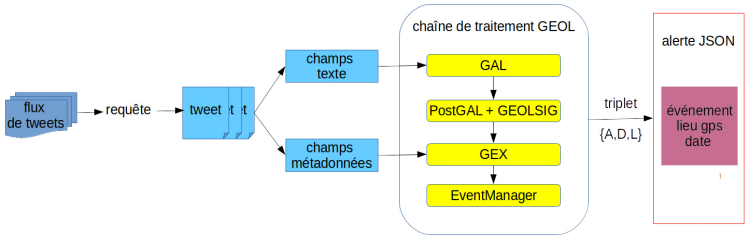

Figure 1. Processus complet de détection d'événements géo-chrono-localisés

3.1. Collecte des tweets

Il existe plusieurs solutions pour collecter des tweets. Citons :

– **l'utilisation de l'API Twitter :** nous pouvons interroger l'API officielle de Twitter pour collecter les tweets et un nombre considérable de métadonnées indisponibles sur l'interface Twitter. Avec la version gratuite, nous ne pouvons récupérer que 1% des tweets à un instant t (par l'API Stream) et filtrer les tweets selon des mots-clés, les coordonnées de géolocalisation et l'identifiant (ID) de l'utilisateur ou la langue de rédaction du tweet. Avec l'API REST, des tweets spécifiques peuvent être extraits, les tweets d'un utilisateur en particulier (les 3200 plus récents), ou par *bounding box* sur un espace géographique donné. Dans ce cas, seuls les tweets des utilisateurs qui autorisent l'application à se connecter à leurs coordonnées GPS sont récupérables. Notons

que dans la version gratuite le nombre de requêtes va de 180 à 450 demandes selon le type d'authentification toutes les 15 minutes [4] (Steinert-Threlkeld, 2018).

– **le *scrapping* :** cette approche consiste à concevoir un programme qui va interroger l'interface web de Twitter afin d'obtenir ensuite la page web de la réponse à la requête et d'en extraire les tweets en relation avec la requête et les métadonnées disponibles, en utilisant l'outil gratuit Twint [5].

Malgré la stabilité et la robustesse de l'API Twitter, compte-tenu des restrictions très fortes quant à son utilisation en version gratuite, nous avons décidé d'utiliser le *scrapping* par Twint car la quantité de tweets est bien supérieure et le texte collecté nous permet de constituer un corpus bien plus complet et pertinent qui sera utilisé par la chaîne de traitement GEOL (*cf.* figure 1).

3.2. Prétraitement des tweets

En plus des dictionnaires argotiques, GEOLSemantics a développé un outil permettant la détection de mots dans un grand flux de données grâce à un dictionnaire arborescent dont chacun des nœuds contient une table de hachage indexant ses branches. Le dictionnaire est d'abord fléchi pour obtenir toutes les formes de chacun des mots. Ce dictionnaire est ensuite phonétisé, puis regénéré en parsant le fichier d'entrée caractère par caractère (*cf.* 2). Le fait de changer les mots de l'argot vers le français et de corriger les fins de mots permet de constituer une entrée qui garantit une analyse sémantique plus propre.

```
Dictionnaire ayant déjà une entrée pour "poire" -> p-o-i-r-e-null
1. On rencontre dans le fichier de génération du dictionnaire le mot "poirier"
2. On suit les branches en cherchant chaque caractère: ROOT-p-o-i-r-
3. On arrive à "i" et on le cherche dans la table de #HashTable_de_poir
4. La table renvoie "null" le i n'existe donc pas, on génère une entrée dans la table pour le i
5. On génère successivement un nœud + une entrée dans ce nouveau nœud pour les caractères suivants
6. On change la valeur de "EndOfWord" a true pour le dernier "r" dans l'arbre poirier
7. On obtient donc l'arbre: ROOT-p-o-i-r-e-null
                                    \
                                     i-e-r-null
```

Figure 2. Processus d'écriture de Poire *et ses déclinaisons dans le dictionnaire*

3.3. Extraction d'événements

Une fois le pré-traitement réalisé, lors de la phase d'extraction, nous avons cherché à repérer des événements composés d'une action, d'un lieu et d'une date. Nous avons vu précédemment que les tweets sont des textes courts, mal rédigés et avares en informations. Aussi, un certain nombre d'adaptations des traitements linguistiques couramment employés sur des documents textuels sont nécessaires. Cette extraction s'appuie principalement sur deux modules (*cf.* GAL et GEX de la figure 1) qui inter-

4. https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets
5. Twitter Intelligence Tool, https://github.com/twintproject/twint/wiki

agissent avec une ontologie propre à GEOLSemantics pour améliorer la sémantique de l'information à traiter.

3.3.1. Ontologie

Cette ontologie se fonde sur des logiques de description qui définissent le sens des concepts par leurs caractéristiques et par une représentation structurée ou formelle de leur rôle dans le domaine. Pour les besoins de nos travaux, nous nous sommes intéressés aux ontologies événementielles. Citons par exemple LODE (Troncy *et al*., 2010), une ontologie pour représenter des événements dans le web des données, SEM qui permet d'identifier les événements passés, présents et futurs dans la presse en relation avec des attaques maritimes (Van Hage *et al*., 2009), *etc.*

Une levée d'ambiguïté est réalisée en confrontant, en particulier, les lieux mentionnés dans le texte avec les lieux trouvés dans les métadonnées et dans les lieux connus du GEOLSIG. L'information ainsi extraite est chargée dans une base indexée par Elasticsearch qui permet de rechercher les entités géographiques semblant les plus pertinentes par rapport à celles repérées dans le texte. Une fois l'interprétation la plus proche choisie, le système retourne les coordonnées de l'entité géographique contenue dans le texte.

3.3.2. Amélioration de la reconnaissance des lieux

Nous avons vu dans l'introduction que dans le cadre du projet *Safecity* nous devons définir la localisation précise d'un événement. GEOLSemantics possède déjà un système de détection des lieux à base d'ontologie complété par des dictionnaires mais qui doit être amélioré. En effet, nos dictionnaires contiennent une liste importante de lieux mais ils ne peuvent pas être exhaustifs. Afin d'augmenter la qualité de la reconnaissance des lieux, nous utilisons des annonceurs de lieux tels que *rivière*, *rue*, *etc*. Cependant, certains lieux non introduits par un annonceur et dont le nom est trop ambigu ne sont pas reconnus par nos analyses. Par exemple, dans la phrase « le loup déborde », seul l'emploi du verbe *déborder* nous indique que *le loup* est un cours d'eau. Le principe est donc d'extraire du tweet une liste de lieux potentiels, en plus des lieux déjà détectés par l'analyse linguistique, puis d'interroger le GEOLSIG pour vérifier si ces lieux existent vraiment. Les nouveaux lieux trouvés sont ensuite réintroduits dans l'analyse afin de mieux reconnaître les événements géo-chrono-localisés.

En utilisant un ensemble de règles linguistiques, nous incluons plus de lieux potentiels, selon la présence ou non d'annonceurs de lieux, de majuscules ou de prépositions particulières. Une interrogation par *Elasticsearch* [6] pour chaque lieu potentiel va permettre de trancher si ces derniers sont effectivement des lieux ou non. Cette phase nous a permis de passer de 14% à 66% de lieux détectés.

Lors de cette interrogation, une requête peut retourner plusieurs solutions. Nous passons alors par une phase de désambiguïsation, en utilisant un système de poids

6. https://www.elastic.co/fr/elasticsearch

permettant de sélectionner le bon résultat parmi les solutions renvoyées. Le poids est la somme des éléments suivants : i) match complet ou partiel entre la requête et le résultat, et vice-versa, avec puis sans les mots vides, ii) catégorie du résultat dans la requête (ex : requête="Aéroport de nice", résultat "nice" catégorie="aéroport"), iii) catégorie présente dans la liste des catégories prioritaires ("ville, commune, quartier, avenue, boulevard, rue, place"), iv) diminution du poids si la catégorie est dans la liste des catégories minoritaires ("arrêt, bus, tram, café"). Le résultat avec le poids le plus important sera renvoyé. Si le poids est négatif aucun lieu ne sera renvoyé.

4. Chaîne de traitement GEOL

Le système d'émission d'alertes à partir des réseaux sociaux développé par GEOLSemantics dans le cadre du projet Safecity consiste en une analyse des tweets émis dans une zone géographique précisée par le client de l'application. Les modules vus précédemment se fondent sur une ontologie et des composants interconnectés présentés dans cette section.

4.1. GAL

Notre système se fonde sur une analyse linguistique profonde (*cf.* module GAL de la figure 1), réalisée grâce à un moteur développé en interne chez GEOLSemantics [7], afin de répondre aux exigences spécifiques à nos besoins. Nous analysons déjà le français, l'anglais et l'arabe. Bien évidemment, chaque nouvelle langue à traiter demande une adaptation des dictionnaires et des règles linguistiques d'extraction sémantique. Les tweets, quant à eux, sont rarement écrits dans un langage soigné mais remplis d'abréviations, sigles, argot, fautes d'orthographe, *etc*. Les premières étapes d'ana-

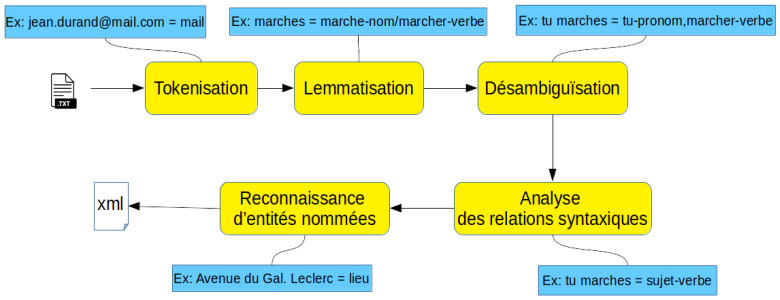

Figure 3. Processus de traitement de l'analyse linguistique profonde

lyse, la tokenisation, la lemmatisation et la désambiguïsation, sont traitées de façon classique mais en les adaptant aux spécificités des tweets.

7. http://www.geolsemantics.com/index.php/fr/

La **tokenisation** permet la reconnaissance des mots-dièse (dits *hashtags*) et des pseudonymes précédés d'une arobase. Cela simplifie l'analyse lorsque certains hashtags sont utilisés comme des mots appartenant à la phrase (par exemple, *"Une habitante se plaint de la présence des dépôts sauvages et des rats à #Courbevoie"*).

Beaucoup de travaux de recherche dans le contexte des tweets (Shoukry, Rafea, 2012) utilisent le *stemming*. Nous avons préféré passer par la **lemmatisation** des mots car nous possédons les ressources nécessaires (dictionnaires de formes fléchies) nous permettant des analyses plus précises. Le traitement des tweets a demandé des adaptations à plusieurs niveaux à savoir : (i) la mise en place d'un *dictionnaire spécialisé* contenant les mots argotiques, les abréviations et les sigles les plus courants, par l'analyse semi-automatique d'un corpus de tweets qui a permis d'intégrer 2179 entrées supplémentaires pour le français ; (ii) le *traitement grammatical contextualisé* des fins phonétiquement identiques : oubli du *s* du pluriel ou de la deuxième personne du singulier des verbes, confusion entre l'infinitif et le participe passé, etc. (iii) le remplacement de la désaccentuation par une *phonétisation* afin de rapprocher les mots lorsqu'ils contiennent des fautes d'orthographe. Par exemple, pour l'analyse de « Gros bazarre sur la #ligneR », après phonétisation le mot « bazarre » aura deux interprétations : *bazar* nom singulier, ou *bazars* nom pluriel ; (iv) le *découpage des mots collés de type hashtags*, en suivant la casse. Par exemple, dans un tweet tel que « La #TableRondeRATP est ouverte », le mot hashtag *#TableRondeRATP* permettra d'obtenir un découpage en *Table Ronde RATP*.

La **désambiguïsation** est effectuée grâce à une méthode statistique fondée sur des trigrammes et des bigrammes de catégories établis à partir d'un corpus étiqueté de tweets. Après la désambiguïsation, nous établissons les **relations syntaxiques** entre les différents éléments de la phrase (nom-adjectif, nom-complément de nom, mais aussi agent-action, action-objet, etc.), et nous détectons les **Entités Nommées.**

L'analyse linguistique nous a permis de lemmatiser et catégoriser les mots des tweets puis d'établir des relations entre les mots afin de simplifier l'étape suivante, c'est-à-dire l'extraction des événements et des informations qui leur sont propres. Aussi, notre système géolocalise les événements par le renforcement de la reconnaissance des entités nommées de lieux en faisant appel à un Système d'Information Géographique (*cf.* GEOLSIG dans la figure 1), créé à partir de fichiers de l'*open data* local et d'une extraction de fichier de l'*open data* global *Openstreetmap* [8].

4.2. PostGAL

L'analyse linguistique de GEOLSemantics (GAL) permet de collecter des noms de lieux. Cet outil repose sur un dictionnaire de noms de lieux connus. Si un nom de lieu est dans une phrase et n'est pas dans le dictionnaire, il est étiqueté en tant qu'Entité Nommée (EN) inconnue (*inc*). Une étape de désambiguïsation s'ensuit afin

8. https://www.openstreetmap.org/

d'attribuer une autre étiquette au mot inconnu (*i.e. loc, pers, org* pour respectivement lieu, personne ou organisation), en fonction de sa position dans la phrase et de ses relations avec les mots l'avoisinant. Par exemple, dans la phrase *Je vais à Kendira*, Kendira est absent du dictionnaire, mais étiqueté *EN loc* grâce au verbe aller. Bien que l'étape de désambiguïsation linguistique permette de collecter plus de lieux, elle reste très stricte en raison du nombre d'erreurs potentielles. Au final, il est préférable d'avoir plus d'*EN inc* que d'EN mal identifiées.

L'utilisation d'un outil de vérification de lieux améliore le taux de désambiguïsation. Avec ce nouveau système, nous interrogeons une base de données de lieux avec les *EN inc*. Si ce nom inconnu est présent dans la base, on peut le réinjecter dans le GAL sous une forme reconnaissable en tant qu'*EN loc*. Pour cela, nous avons établi une liste de règles linguistiques qui augmente la permissivité, et permet à plus de noms de lieux probables d'être collectés. Ces règles vont également permettre de compléter les noms de lieux et de potentiellement les regrouper avec des éléments les entourant. Par exemple, *Je suis au MK 2 de Bercy* devient *Je suis au MK 2 de Bercy*; après complétion, *MK* est réuni avec *2*, puis *MK 2* avec *de Bercy*). En outre, beaucoup de lieux intègrent des noms de personnes. Là encore nous pouvons utiliser des règles linguistiques pour vérifier si une *EN pers* détectée n'est pas, en fait, un lieu. Il faut s'assurer que tous les composants du nom propre sont présents dans le lieu retenu. Par exemple, si l'on recherche *Nicolas Sarkozy*, il faut que *Nicolas* et *Sarkozy* soient présents dans le résultat.

La figure 4 présente l'architecture de ce système. Il s'agit de l'ensemble des traitements effectués après l'analyse linguistique, d'où le nom postGAL.

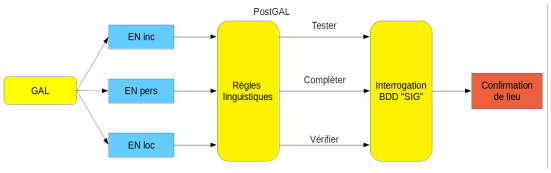

Figure 4. Organisation du traitement PostGAL

4.3. GEOLSIG

Comme mentionné en section 4.5, les événements découverts (grâce au module GEX) sont insérés dans une base et indexés par Elasticsearch. Ils y sont représentés au format JSON suivant une hiérarchie établie, à laquelle nous rajoutons plusieurs champs de dates. L'insertion dans Elasticsearch nous permet par la suite d'exécuter les requêtes pour la fusion d'événements, qu'il s'agisse de collecter, ajouter ou supprimer des entrées. Le contenu de la base est aussi utilisé pour la visualisation finale ou soumis à d'autres modifications. Un nettoyage de la BD est réalisé régulièrement dans le but de supprimer les événements (dits périmés) dont la date de validité est dépassée. Ainsi, la base conserve une taille réduite, assurant de meilleures performances pour les divers programmes l'utilisant. Par soucis d'anonymat, chaque événement est

caractérisé par un identifiant unique attribué avant son insertion dans la base et généré par nos soins.

4.4. GEX

Notre extraction (*cf.* module GEX dans la figure 1) permet de repérer des entités nommées (personnes, organisation, lieu, date, etc.) et des événements (incendie, accident, etc.) pouvant nécessiter l'intervention des autorités locales (pompiers, ambulance, police). Le GEX utilise des motifs syntactico-sémantiques guidés par une ontologie. Nous avons élaboré une ontologie permettant la modélisation des événements repérés dans les textes. Cette ontologie (*cf.* tableau 1) a été créée à l'aide de Protégé [9] au regard des spécifications du système d'alerte puis des différents corpus des événements. Cette ontologie fait partie d'une ontologie plus générale de la société GEOLSemantics, dans laquelle nous avons créé 170 classes, 363 ObjectProperties, 181 DatatypeProperties et déjà développé les définitions des entités nommées (de type lieu, date, *cf.* tableau 2.) relatives aux connaissances communes à divers domaines. [10].

Tableau 1. Ontologie partielle des événements

Concept	Sous-type d'action	Lieu	Date	Entité concernée
Incendie	event-type	event-place	event-date	isBurned
Accident	event-type	event-place	event-date	involve
Inondation	event-type	isFlooded	event-date	inundationSource
Colis suspect	package-type	event-place	event-date	–

Tableau 2. Ontologie partielle pour les entités

Concept	Propriétés
Location	loc-type, loc-name, locatedIn, hasGeoPosition
GeoPosition	coordinates
Date	dateBeg, dateEnd

Les patrons syntactico-sémantiques (dit pattern d'extraction) sont créés afin d'extraire les informations à partir des résultats de l'analyse linguistique profonde. Un pattern d'extraction est composé de trois éléments : i) un concept et une propriété définis par l'ontologie et permettant la représentation de l'information extraite au format RDF (Cyganiak *et al*., 2014) ; ii) un patron lexico-syntaxique qui contient les informations morpho-syntaxiques d'une relation syntaxique ; iii) et un contexte constitué d'une ou plusieurs relations syntaxiques, aidant à la désambiguïsation de sens de l'action et de type de l'entité. Le GEX propose également une résolution des dates, consistant à transformer les dates relatives identifiées au niveau du module GAL comme *demain* ou *hier* en dates absolues en utilisant une date de référence. Il s'agit de la date absolue

9. voir https://protege.stanford.edu/

10. Cette ontologie n'est pas disponible pour des raisons commerciales

citée dans le texte ou, en cas d'absence de celle-ci, la date d'émission présente dans les métadonnées du tweet. Pour le cas particulier des tweets, lorsque les émetteurs du message ne précisent aucune date et que les verbes indiquent un événement en cours, alors nous attribuons à l'événement la date d'émission du tweet.

4.5. Event manager

Il est possible que des événements extraits soient similaires. Pour cette raison, chaque événement nouvellement extrait doit être comparé avec ceux précédemment collectés. S'ils s'avèrent similaires, ils sont alors fusionnés. La fusion est effectuée lorsqu'il y a une intersection temporelle et spatiale entre les événements, et que les actions qu'ils représentent sont identiques. Afin de vérifier une telle similarité, nous procédons en deux temps. Une requête permet de récupérer dans notre base de données les événements ayant des actions identiques et dont les intervalles de validité se chevauchent, puis l'intersection spatiale est calculée en utilisant une bibliothèque Java appropriée. Nous aurions également pu ajouter une clause à notre requête pour vérifier ce dernier point, mais le système d'indexation Elasticsearch [11] que nous utilisons ne permet pas de calculer l'intersection entre les différents types de surfaces, chose nécessaire pour la représentation fusionnée.

Lorsqu'ils sont regroupés, les événements conservent la liste des tweets qu'ils représentent. Les dates et les lieux doivent alors être précisés: pour les dates de début et de fin d'événement, nous considérons les valeurs les plus anciennes comme étant les plus précises. En ce qui concerne les lieux, l'intersection des représentations est adoptée, et dans le cas d'événements localisés par des points, nous effectuons une vérification de distance à la place de l'intersection, en utilisant une valeur de référence paramétrable. Nous considérons qu'un point est toujours jugé plus précis par rapport à un polygone ou un multipoint. Pour la comparaison d'événements représentés chacun par un point, si la distance les séparant est inférieure à la valeur de référence, nous considérons qu'il y a intersection, et utilisons le centre du segment formé par les deux points comme localisation dans la fusion des événements.

Lors de la fusion, le reste des métadonnées de l'événement généré provient de l'événement le plus ancien. Néanmoins, puisque chaque événement contient la liste des tweets qui l'a généré, l'ensemble des informations de base reste disponible dans le résultat final pour tout utilisateur. Nous avons appelé le composant responsable de ce traitement l'*Event Manager* (*cf.* figure 1).

11. voir https://www.elastic.co/fr/Elasticsearch

5. Évaluations

5.1. Expérimentations

BERT [12] (Devlin *et al*., 2019) est un modèle contextuel de langues, fondé sur les réseaux de neurones, développé chez Google et diffusé en *open source* en 2018. Le succès révolutionnaire de BERT et de la classe de modèles appelés "transformeurs" (BERT, ALBERT, RoBERTa, GPT, GPT-2, etc.), a permis de mettre au point des modèles linguistiques capables d'atteindre des performances record dans les différentes tâches du NLU "Natural Language Understanding". CamemBERT et FlauBERT sont deux variantes qui offrent les meilleurs résultats pour la langue française. L'avantage d'utiliser ces modèles est la réduction du coût de préparation du corpus d'apprentissage, car ils sont pré-entraînés sur des corpus très volumineux en français issus de différentes sources (Wikipedia, livres, internet, journaux, etc.). Notons que le plongement contextuel (*Contextual embeddings*), comme BERT, va au-delà des représentations des plongements de mots (*Word embedding*), comme Word2Vec, et atteint des performances révolutionnaires sur un large éventail de tâches de traitement du langage naturel puisqu'il attribue à chaque mot une représentation basée sur son contexte. (Liu *et al*., 2020) ont examiné les modèles existants d'intégration contextuelle.

Pour la comparaison avec notre système, nous avons affiné le réseau de neurones fourni par CamemBERT en lui ajoutant une couche de sortie concernant la reconnaissance des deux entités "ACTION" et "LOCATION". Nous avons entraîné ce nouveau modèle sur un corpus d'apprentissage composé de 1195 tweets sur les inondations dans la région de la Côte d'Azur en France, sur une période allant de 2012 jusqu'à 2019. Le corpus intégral est composé de 1495 tweets, annotés en précisant le lieu, le déclencheur de l'action *inondation* (inondation, montée des eaux, etc.), et si le tweet représente un événement ou non. 300 tweets ont été isolés après annotation et utilisés comme corpus de test pour les deux systèmes.

Pour notre solution, nous avons pris un corpus de développement de 118 tweets (10%) issus du corpus d'apprentissage de BERT pour affiner nos règles. Après l'affinage des règles sur les deux systèmes à comparer, nous avons lancé les analyses sur le corpus de test restant de 300 tweets. Les résultats sont présentés dans le tableau 3.

Afin de valider les fusions d'événements, plusieurs des tweets utilisés présentaient des similarités partielles ou totales (considérant l'action et les localisations temporelles et géographiques). Ceci nous a permis de confirmer le bon fonctionnement de la fusion de tweets. Les résultats ont également pu être analysés afin de s'assurer que l'intégralité des champs JSON du regroupement - identifiant, dates, intersection - étaient bien remplis et au bon format.

12. Bidirectional Encoder Representations from Transformers

5.2. Analyse des résultats

Dans un premier temps, nous avons remarqué que BERT est efficace pour détecter des lieux et des actions mais présente des lacunes en sémantique. La F-mesure de la détection des actions seules (resp. lieux) est de 88% contre 70% pour GEOLSemantic (resp. 96% contre 88%). En effet, un message peut inclure un déclencheur d'action et un lieu sans pour autant représenter un événement (négation, événement très ancien, métaphore, etc.), tandis que notre système a une plus grande précision (88%) et ne renvoie que les événements sémantiquement corrects. Par exemple, *"Mon balcon est inondé, c'est une piscine olympique le truc"* est renvoyé par BERT, mais pas par GEOLSemantics.

Puisque notre outil est plus correct dans son traitement des événements qui ne doivent pas déclencher d'alerte, nous avons pris les résultats de GEOLSemantics (GEOL) sur les non-événements et les résultats de BERT sur les événements, constituant ainsi un système hybride qui arrive à 89% de F-mesure tout en étant meilleur en rappel et en précision que chacun des systèmes exécutés individuellement. Enfin nous avons étudié les cas où au moins l'un des systèmes avait une bonne réponse. Cette approche a obtenu 95% de F-mesure. Il semble que les deux systèmes soient complémentaires et qu'il faille les configurer selon le contexte. Concernant les temps d'exécution, BERT est plus rapide, ce qui s'explique par la quantité de traitements que nous effectuons. Nous pallions ceci en effectuant de la parallélisation des traitements à tous les niveaux d'analyse, mais aussi en ajoutant une première étape qui applique un filtre « à grandes mailles » sur le flux des tweets d'entrée, afin d'analyser que ceux qui ont une chance de contenir des informations pertinentes. Ce filtre s'appuie sur une liste de mots-clés pouvant représenter un événement à extraire.

Tableau 3. Tests de performance des systèmes GEOLSemantics et BERT

	GEOL	Bert	Non événement–>GEOL évenement–>BERT	Combinaison GEOL/BERT
Rappel	86%	90%	90%	99%
Présision	88%	74%	89%	91%
F-mesure	87%	80%	89%	95%

6. Conclusion et perspectives

Nous avons montré qu'un processus de traitement complet combinant des approches linguistique et sémantique afin de compléter les informations composant un événement par recoupement et regroupement est possible. Un système hybride nous permet d'obtenir des résultats très satisfaisants pour une première expérience, résultats améliorables en configurant mieux BERT et en rendant notre système encore plus robuste.

Diverses améliorations sont planifiées pour l'algorithme de regroupement d'événements ayant un contexte commun partiel. Par exemple, regrouper des événements de

localisation géographique et de temporalité similaires mais dont l'action représentée est différente. En effet, malgré cette divergence d'action, les autres éléments détectés laissent suggérer que de tels événements sont liés d'une manière ou d'une autre. D'autre part, il faut affiner la modélisation des zones géographiques des événements à fusionner. En effet, dans plusieurs cas, utiliser l'intersection entre chacun des lieux déterminés s'avère moins pertinent que de conserver une zone trouvée dans un des événements à fusionner.

Dans cet article, nous avons expérimenté notre système sur un corpus d'inondation. Pour nos travaux futurs, nous souhaitons étudier la pertinence de notre système dans d'autres contextes. En effet, le tableau 3 a montré que d'autres contextes peuvent influencer le choix du système. Un nouveau corpus (7627 tweets sur des incendies sur la Côte d'Azur pour une période allant de 2012 à 2020) va également nous permettre d'étudier les capacités de BERT pour la détection des multi-événements, tâche déjà développée dans notre système. Nous allons aussi tester la rapidité et la précision des deux systèmes en simulant un flux de tweets en temps réel comportant beaucoup de bruit (tweets non pertinents).

Remerciements

Cet article est dans le cadre du projet Safecity soutenu par la BPI. Nous remercions Houda Saadane, El Mehdi Khalfi et Christian Fluhr pour leur implication.

Bibliographie

Aramaki E., Maskawa S., Morita M. (2011). Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, p. 1568–1576.

Besnard P., Cordier M.-O., Moinard Y. (2008). Ontology-based inference for causal explanation. *Integrated Computer-Aided Engineering*, vol. 15, n° 4, p. 351–367.

Broniatowski D. A., Paul M. J., Dredze M. (2013). National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one*, vol. 8, n° 12.

Cyganiak R., Wood D., Lanthaler M. (Eds.). (2014). *Rdf 1.1 concepts and abstract syntax.* Consulté sur https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

Devlin J., Chang M.-W., Kenton L., Toutanova K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In . Consulté sur https://arxiv.org/pdf/1810.04805.pdf

Guibon G., Ermakova L., Seffih H., Firsov A., Le Noé-Bienvenu G. (2019). Multilingual fake news detection with satire.

Han B., Cook P., Baldwin T. (2013). A stacking-based approach to twitter user geolocation prediction. In *Proceedings of the 51st annual meeting of the association for computational linguistics: system demonstrations*, p. 7–12.

Hecht B., Hong L., Suh B., Chi E. H. (2011). Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the sigchi conference on human factors in computing systems*, p. 237–246.

Hoang T. B. N., Mothe J. (2018). Location extraction from tweets. *Information Processing & Management*, vol. 54, n° 2, p. 129–144.

Huang B., Carley K. M. (2019). A hierarchical location prediction neural network for twitter user geolocation. *arXiv preprint arXiv:1910.12941*.

Jackoway A., Samet H., Sankaranarayanan J. (2011). Identification of live news events using twitter. In *Proceedings of the 3rd acm sigspatial international workshop on location-based social networks*, p. 25–32. New York, NY, USA, Association for Computing Machinery. Consulté sur https://doi.org/10.1145/2063212.2063224

Khamis S., Ang L., Welling R. (2017). Self-branding,'micro-celebrity'and the rise of social media influencers. *Celebrity studies*, vol. 8, n° 2, p. 191–208.

Kleinberg J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, vol. 7, n° 4, p. 373–397.

Kryvasheyeu Y., Chen H., Moro E., Van Hentenryck P., Cebrian M. (2015). Performance of social network sensors during hurricane sandy. *PLoS one*, vol. 10, n° 2.

Liu Q., Kusner M. J., Blunsom P. (2020). A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.

Perez S. (2017, Nov). *Twitter officially expands its character count to 280 starting today.* TechCrunch. Consulté sur https://techcrunch.com/2017/11/07/twitter-officially-expands-its-character-count-to-280-starting-today/?guccounter=1

Sakaki T., Okazaki M., Matsuo Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web*, p. 851–860.

Shoukry A., Rafea A. (2012). Preprocessing egyptian dialect tweets for sentiment mining. In *The fourth workshop on computational approaches to arabic script-based languages*, p. 47.

Steinert-Threlkeld Z. C. (2018). *Twitter as data*. Cambridge University Press. Consulté sur https://www.cambridge.org/core/elements/twitter-as-data/27B3DE20C22E12E162BFB173C5EB2592

Troncy R., Shaw R., Hardman L. (2010). Lode: une ontologie pour représenter des événements dans le web de données. In *21es journées francophones d'ingénierie des connaissances (ic 2010), http://www. ic2010. mines-ales. fr/*, p. 69–80.

Van Hage W. R., Malaisé V., Vries G. de, Schreiber G., Someren M. van. (2009). Combining ship trajectories and semantics with the simple event model (sem). In *Proceedings of the 1st acm international workshop on events in multimedia*, p. 73–80.

Wang X., Gerber M. S., Brown D. E. (2012). Automatic crime prediction using events extracted from twitter posts. In *International conference on social computing, behavioral-cultural modeling, and prediction*, p. 231–238.

Weng J., Lee B. (2011, 01). Event detection in twitter.

Ying Y., Peng C., Dong C., Li Y., Feng Y. (2018). Inferring event geolocation based on twitter. In *Proceedings of the 10th international conference on internet multimedia computing and service.* New York, NY, USA, Association for Computing Machinery. Consulté sur https://doi.org/10.1145/3240876.3240909