DataNews : Contextualisation de valeurs chiffrées dans des dépêches

DataNews: contextualisation of quantified values in wires

Chloé Monnin[1], Olivier Hamon[1], Victor Schmitt[2], Brice Terdjman[2]

[1] Syllabs, Paris, France
[2] WeDoData, Paris, France

RÉSUMÉ. L'Open Data fournit de nombreuses données publiques avec une couverture très large, mais aucune base n'a jamais été structurée à partir d'informations issues de l'actualité. À travers DataNews, notre objectif est d'aller chercher automatiquement des données afin d'offrir un moyen de les réutiliser. Pour ce faire, nous avons tout d'abord défini une typologie d'événements dans le contexte spécifique des décès dans des dépêches AFP. Puis, en se limitant aux catastrophes naturelles, nous avons regroupé ces dépêches par événement afin de pouvoir les identifier. La dernière étape a pour objectif de construire des patrons d'extraction afin de collecter les valeurs correspondant au nombre de morts, de même que le contexte associé à ces valeurs. Les résultats de nos évaluations nous ont confirmé le fort potentiel de notre méthode qui pourrait amener à l'élaboration de plusieurs applications.

ABSTRACT. The Open Data allows the access to plentiful data, with a large coverage, but none of them offers a structured databased around news. Through DataNews, our goal is to seek for data automatically so as to provide means to reuse them. To do so, we first defined an event typology in the specific context of death in AFP wires. Then, by restraining ourselves to the natural disasters, we clustered these wires by events so as to identify them. The goal of the last step is to build extraction patterns so as to collect values corresponding to the death number, as well as the context associated to these values. The results of our evaluations reassured ourselves in the large potential of our method that could lead to several applications.

MOTS-CLÉS. base de connaissances, extraction d'information, construction de patrons, détection d'événements.

KEYWORDS. knowledge base, information extraction, pattern building, event detection.

1. Introduction

Le projet DataNews[1] a pour objectif de construire automatiquement une base de connaissances permettant de capitaliser l'information sur les valeurs la constituant. Il créera ainsi des bases essentielles pour retracer l'historique d'une donnée et donc permettre de comparer l'information, pour détecter des fake news ou pour enrichir les angles journalistiques et les perspectives des lecteurs. Pour ce faire, des valeurs chiffrées sont extraites automatiquement depuis des dépêches catégorisées au préalable. Dans le cadre de ce projet, nous avons utilisé un corpus de dépêches AFP (Agence France Presse). L'intérêt d'extraire des valeurs à partir de dépêches AFP est triple, puisque les dépêches sont précises, factuelles et impartiales. La 3e agence de presse mondiale est un support médiatique de confiance comptant plus de 2 500 journalistes sur les 5 continents qui est utilisé par les principaux médias mondiaux quotidiennement, avec un volume de données et une couverture inégalables : chaque heure, l'agence produit 208 dépêches, 10 vidéos, 125 photos et 3 infographies. L'ensemble du processus de DataNews est présenté en Figure 1.

La base de connaissances que l'on cherche à constituer se veut unique et pratique pour les journalistes puisque les données qu'elle contiendrait sont, de par leur nature, spécifiques, dispersées et rarement rencontrées dans d'autres bases de données. De plus, cette base de connaissances n'a pas pour vocation de ne contenir les seules valeurs chiffrées, mais également leur contexte. Nous avons concentré nos travaux sur un périmètre bien réduit afin de tester nos méthodes dans un premier temps. Ce périmètre concerne le nombre de décès lors d'événements, quels qu'ils soient. Notre objectif est d'étendre ce périmètre peu à peu vers d'autres types d'informations chiffrées, telles que des valeurs monétaires, le nombre de participants à des manifestations, etc.

[1] Projet Digital News Innovation Fund (DNI Fund) DataNews, DNI-r4-6PXRlnq3dAow

Deux partenaires viennent constituer le projet : WeDoData, une agence de datajournalisme à l'initiative du projet, experte en design d'informations et qui a collaboré à la définition de la typologie, a réalisé les validations et a pour objectif d'utiliser les données extraites pour de la dataviz ou de l'aide à la rédaction journalistique ; Syllabs, spécialisée en sémantique et qui a travaillé sur la collecte des données, la définition de la typologie, la définition des patrons, l'extraction des valeurs et leur restitution dans une base de connaissances.

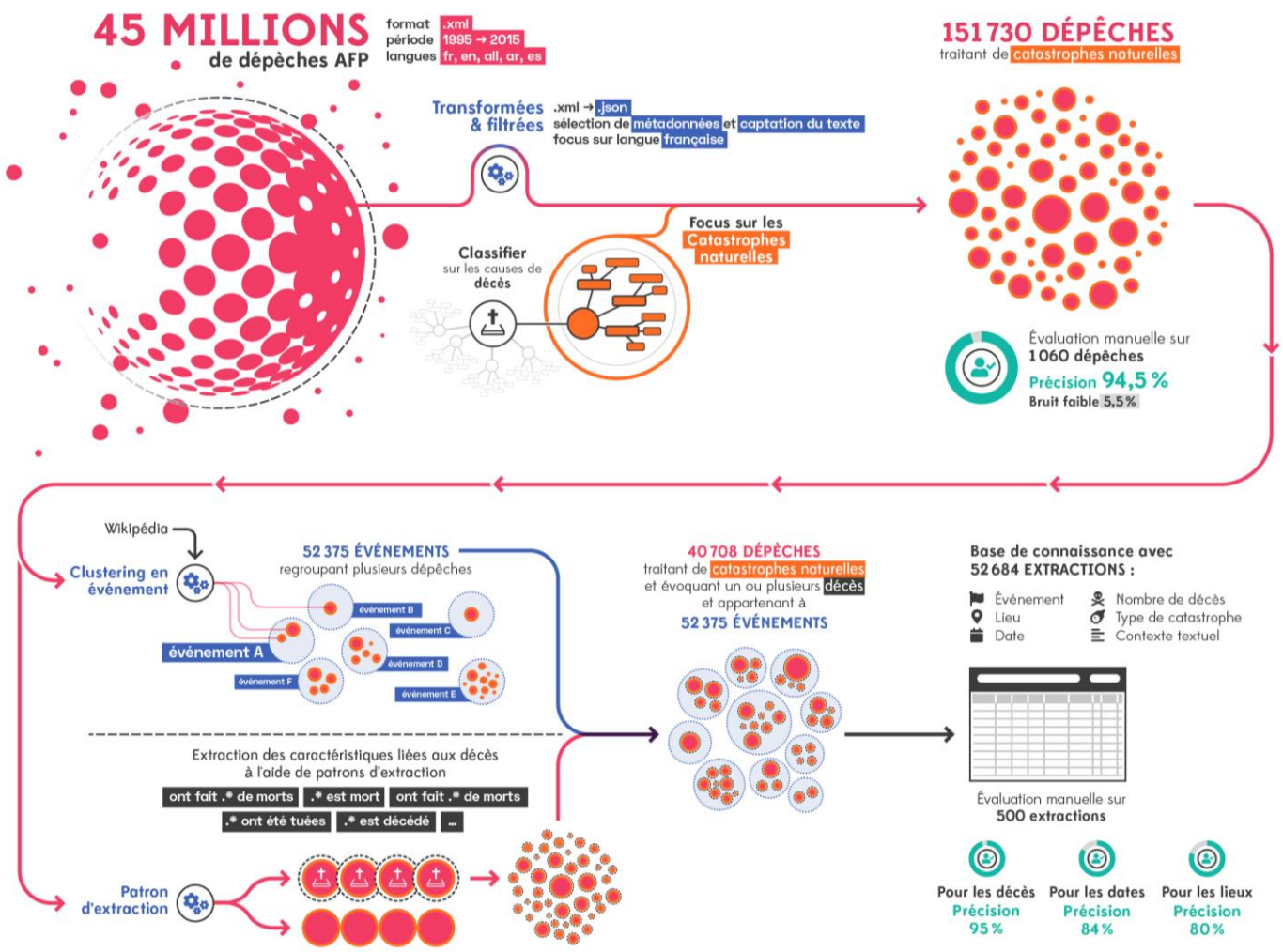

Figure 1. *Processus global de DataNews.*

Après un état de l'art sur la définition de patrons et l'extraction de valeurs, et une contextualisation des travaux menés nous organisons cet article de la même manière que les étapes clés du projet : constitution des données, définition d'une typologie d'événements, extraction de valeurs et évaluation des résultats.

2. État de l'art

L'extraction d'informations a toujours été un domaine prédominant dans le traitement automatique du langage. Cette tâche consiste en la détection et l'extraction sous forme structurée, de données présentes dans des documents non structurés. Le développement des méthodes d'extraction a été stimulé par les conférences MUC (*Message Understanding conferences*) dès la fin des années 80, en organisant des compétitions financées par la DARPA [GRI 96]. L'objectif de ces conférences était d'extraire le plus d'informations possible sur des thèmes bien déterminés et d'évaluer les systèmes d'extraction d'informations selon une grille d'évaluation commune. En dehors de ces conférences, la communauté scientifique a développé d'autres méthodes, utilisant notamment l'extraction de relations à l'aide de patrons.

[BRI 98] avec son système DIPRE (*Dual Iterative Pattern Relation Extraction*) a proposé une méthode permettant d'exploiter le potentiel de la multiplicité de sources d'informations présentes sur le web. Cette méthode génère des patrons à partir d'exemples de valeurs fournies en entrée par l'utilisateur. Les patrons extraits sont ensuite utilisés pour rechercher de nouvelles instances de la relation que le patron représente. Contrairement à DIPRE, les méthodes suivantes sont semi-supervisées, puisque leur point d'entrée n'est pas une valeur mais un corpus. En effet, contrairement aux méthodes ci-dessus, celles-ci ont pour but d'extraire toutes les relations au sein d'un corpus défini, et non une relation particulière servant d'amorce.

[LIN 01] font l'hypothèse distributionnelle pour développer DIRT, dont le but est d'extraire des relations d'inférence (dans ce cas précis des paraphrases). Leur méthode permet d'extraire ce qu'ils nomment des *slot fillers*, soit des entités qui apparaissent dans les trous laissés par leur patrons à partir d'arbres syntaxiques en dépendances, de règles de comparaison et d'un calcul de similarité afin d'identifier les paraphrases. [ETZ 2004] publient les premiers résultats de leur système KnowItAll, un outil permettant d'extraire ce qu'ils nomment des faits, soit des informations mises en relations. Leurs travaux se poursuivront les années suivantes avec KnowItNow [CAF 05]. Leur méthode s'appuie sur des clauses, soient des règles prédéfinies pour extraire des entités qui répondent à ces règles. Les améliorations de KnowItNow concernent principalement la rapidité d'exécution de l'outil ainsi que l'apprentissage de règles d'extraction. [ETZ 11] approfondissent leur système et développent Reverb, un système d'extraction de relations basées sur les verbes. Ce système applique un filtre syntaxique reposant sur des patrons morphosyntaxiques (des expressions régulières de parties du discours) et un filtre lexical. Plus récemment, [AKB 12] ont étendu l'extraction de relation non supervisée en extrayant toutes les relations d'un corpus puis en les classant par *clustering*.

Les travaux cités ci-dessus sont implémentés pour des contenus en anglais. Pour un contexte français, il est possible de trouver des méthodes similaires appliquées à des documents techniques, comme par exemple le système Prométhée [MOR 99] qui extrait des schémas lexico-syntaxiques représentatifs d'une relation sémantique. Par ailleurs, concernant la détection d'événements sur des corpus journalistiques pour de la mise en contexte d'informations précises, nous pouvons citer [BOS 08] ainsi que [RIB 17] qui utilisent le *clustering* pour rassembler des documents mentionnant les mêmes événements. À terme, notre objectif se rapproche de celui poursuivi par *L'Europe Media Monitor*, une base de données automatiquement alimentée par les articles en lignes et réseaux sociaux européens qui identifie des événements en temps réel[2], ainsi que des travaux de Ritchie et Roser[3].

3. Contexte

L'Open Data fournit de nombreuses données publiques avec une couverture très large, mais aucune base n'a jamais été structurée à partir d'informations issues de l'actualité. Il n'est pas rare de voir des projets utilisant des approches semi-supervisées ou non supervisées pour constituer des bases de données à partir de corpus textuels. Mais, à notre connaissance, aucun n'a été dédié à créer une base de connaissances pour des salles de rédaction.

Que ce soit le nombre de morts lors d'un séisme ou la somme dépensée pour acheter un joueur de foot, une quantité énorme de données, potentiellement chiffrées, sont créées tous les jours dans les flux de dépêches. Ces chiffres restent habituellement au sein d'un texte en langage naturel et ne sont pas stockés dans une base de données structurées, et ne sont donc pas capitalisées. Alors qu'ils sont indispensables pour la compréhension de l'actualité, ils disparaissent dans le flot énorme de dépêches produites. De plus, pour des raisons de crédibilité,

[2] https://ec.europa.eu/jrc/en/scientific-tool/europe-media-monitor-newsbrief

[3] http://ourworldindata.org

il est essentiel pour les médias de fournir les bons chiffres, les bonnes comparaisons à leurs lecteurs. Notre projet a pour vocation de donner de la perspective aux données chiffrées. Par exemple, si un journaliste présente les revenus annuels de l'industrie de la bière comme étant 523 milliards de dollars, il est difficile de se faire une idée raisonnable du montant. En revanche, si le journaliste est capable de dire que cette valeur correspond au budget spatial américain de 1959 à 2005, cela donne un sens à cette information, l'illustre d'une manière plus cohérente.

Notre objectif est ainsi d'aller chercher automatiquement ces valeurs afin d'offrir un moyen de les réutiliser. Par ailleurs, cela ne peut se faire qu'en contextualisant les valeurs collectées, en extrayant également les unités, les dates, les lieux, des catégories, etc. Le sujet, l'événement lié à une valeur donnée est également de toute importance, ce qui permettra notamment de suivre l'évolution des valeurs au cours du temps.

4. Constitution des données

L'élément indispensable à la réalisation de notre projet résidait dans l'acquisition d'un corpus suffisamment large, et donc susceptible de contenir des volumes importants de données structurées. Plus il y a de données, plus nous augmentons les possibilités de collecter un maximum de données, même dans un périmètre restreint.

Nous sommes partis d'un très grand corpus de dépêches AFP écrites sur 20 ans (1995 – 2015), l'intérêt principal étant de pouvoir disposer d'un important corpus d'entraînement mais aussi, dans une moindre mesure de couvrir une large période et ainsi accroître la possibilité d'obtenir des événements correspondant à notre périmètre. C'est aussi un moyen d'obtenir différents types de dépêches, écrites par différentes personnes, dans différents contextes et sur plusieurs années. Le contexte bien réel de ces contenus a également toute son importance, car il permet d'obtenir des formes riches et variées.

Les données d'origines sont au format XML d'après le standard NewsML-G2[4] et contiennent un grand nombre de métadonnées dont, toutefois, la présence et la qualité sont variables. Nous avons transformé ces données dans un format, une structure, et avec des métadonnées plus conformes à nos besoins. Outre la facilité d'accès, l'objectif était ici de réduire le volume des données à notre contexte d'utilisation et ainsi faciliter les traitements et leur observation.

Le format choisi a été le JSON puisqu'il est utilisé dans tous nos traitements internes, et la sélection des champs s'est faite selon les métadonnées potentiellement utiles aux traitements ultérieurs, à savoir :

- *<sent>/date_time* : l'heure et la date d'envoi de la dépêche ;
- *<located>/location* : le lieu où se déroule l'événement rapporté ;
- *<headline>/title* : le titre de la dépêche ;
- *<keyword>/keywords* : les mots clés enrichissant la dépêche ;
- *<language>language* : le code de la langue dans laquelle est rédigée la dépêche
- *<contentSet>text* : le texte de la dépêche ;
- *<urgency>/urgency* : le niveau de priorité de la dépêche, de 1 (plus prioritaire) à 4 (moins prioritaire) ;
- *<creator>/creator* : le nom et le rôle de l'auteur (ou des auteurs) de la dépêche, si présent ;
- *<subject>/event_id* : l'identifiant du thème d'après le standard IPTC[5], si présent.

La Figure 2 présente un exemple de dépêche obtenue et la structure adoptée.

L'ensemble des données ainsi formatées ont été exportées dans une base de données afin de faciliter les accès et échanges. Nous avons découpé le corpus afin d'obtenir un échantillon sur lequel réaliser nos

[4] https://www.afp.com/communication/iris/Guide_to_AFP_NewsML-G2.html

[5] http://cv.iptc.org/newscodes/mediatopic/

entraînements et tests. En premier lieu, nous nous sommes limités à l'année 2005 du fait des catastrophes naturelles qui ont eu lieu cette année, comme les effets du tsunami du 26 décembre 2004 ou la tempête Katrina en août 2005. Nous nous sommes également limités aux dépêches en français, quand bien même notre méthode peut s'appliquer à d'autres langues. Par la suite, nous avons utilisé notre classification automatique d'événements (voir ci-dessous) sur ces dépêches pour ne garder que les dépêches mentionnant des catastrophes naturelles. Enfin, nous avons supprimé du corpus les dépêches dites « multi-sujets », telles que les revues de presse ou les rappels de titres, à partir de mots-clés contenus dans les titres.

La méthode a ensuite été reportée sur l'ensemble du corpus de plus de 37 millions de dépêches et, au final, 11 millions dépêches en français ont été sélectionnées, puis 151 000 mentionnant des catastrophes naturelles. Une dépêche peut aller d'un texte d'une ou deux phrases, à un document de plusieurs milliers de termes, la plupart étant composées de plusieurs paragraphes. La taille dépend généralement de l'évolution d'un fait.


```
{
  "creator": {
    "name": "PHILIPPE MERLE",
    "role": "afpcrrol:photographer afpctrol:forbyline"
  },
  "date_time": {
    "date": "2013-11-17",
    "time": "06:11:41"
  },
  "event_id": [
    "subj:03013000",
    "medtop:20000139"
  ],
  "keywords": [
    "ACCIDENT",
    "FRANCE",
    "MAROC",
    "TOURISME"
  ],
  "lang": "fr",
  "location": "Lyon Rhône FRANCE",
  "text": "Etienne Belot (G) Pierre Cohu (D) et les autre des rescapés d'un raid dans l'Atlas marocain, répondent aux questions des journalistes, le 21 septembre 2005, à l'aéroport Saint-Exupery de Lyon. Ces quatre personnes rentrées en France étaient sorties indemnes de la tempête de neige survenue le 17 septembre sur le mont M'goun qui culmine à 4.068 mètres près de Ouarzazate (550 km au sud-est de Rabat).
AFP PHOTO PHILIPPE MERLE
Etienne Belot (G) Pierre Cohu (D) et les autre des rescapés d'un raid dans l'Atlas marocain, répondent aux questions des journalistes, le 21 septembre 2005, à l'aéroport Saint-Exupery de Lyon. Ces quatre personnes rentrées en France étaient sorties indemnes de la tempête de neige survenue le 17 septembre sur le mont M'goun qui culmine à 4.068 mètres près de Ouarzazate (550 km au sud-est de Rabat).
AFP PHOTO PHILIPPE MERLE",
  "title": "MAROC-FRANCE-TOURISME-ACCIDENT",
  "urgency": "4"
}
```


Figure 2. *Exemple de dépêche au format JSON.*

5. Typage d'événements

5.1. *Typologie d'événements liés au décès*

Nous avons défini une typologie d'événements pour catégoriser les dépêches en rapport avec des décès et, par la même occasion, relier les valeurs extraites à ces types d'événements. Il n'existe pas à notre connaissance de telle topologie, tout du moins conforme à nos attentes. Notre typologie a donc été construite à partir d'observations du corpus et d'articles de presse (utilisant le mot-clé « mort ») et de recherches sur Internet

(Wikipédia, rapports de l'OMS, etc.). Elle est hiérarchique et représentée sous forme d'un arbre dont chaque feuille correspond à une catégorie de cause de décès. La Figure 3 montre une version graphique de la typologie restreinte aux événements liés à des décès lors de catastrophes naturelles.

A partir de nos observations, nous avons identifié cinq grands types de causes de décès qui se divisent selon des sous-types sur plusieurs niveaux, à savoir :

– guerres : morts militaires (par exemple *morts au front*), morts civiles (par exemple *pertes civiles*), morts collatérales (par exemple *déplacements*) ;

– morts intentionnelles : attentats (par exemple *voiture piégée*), meurtres/assassinats (par exemple *fait divers*), morts politiques (par exemple *génocide*), suicides ;

– accidents : maritimes (par exemple *naufrage*), aériens (par exemple *crash d'avion*), routiers (par exemple *accident de la route*), autres causes accidentelles (par exemple *incendie*) ;

– maladies : contagieuses (par exemple *épidémie*), non contagieuse (par exemple *cancer*) ;

– catastrophes naturelles : séisme, glissement de terrain, avalanche, canicule, tsunami, éruption volcanique, inondation, tempête, incendie.

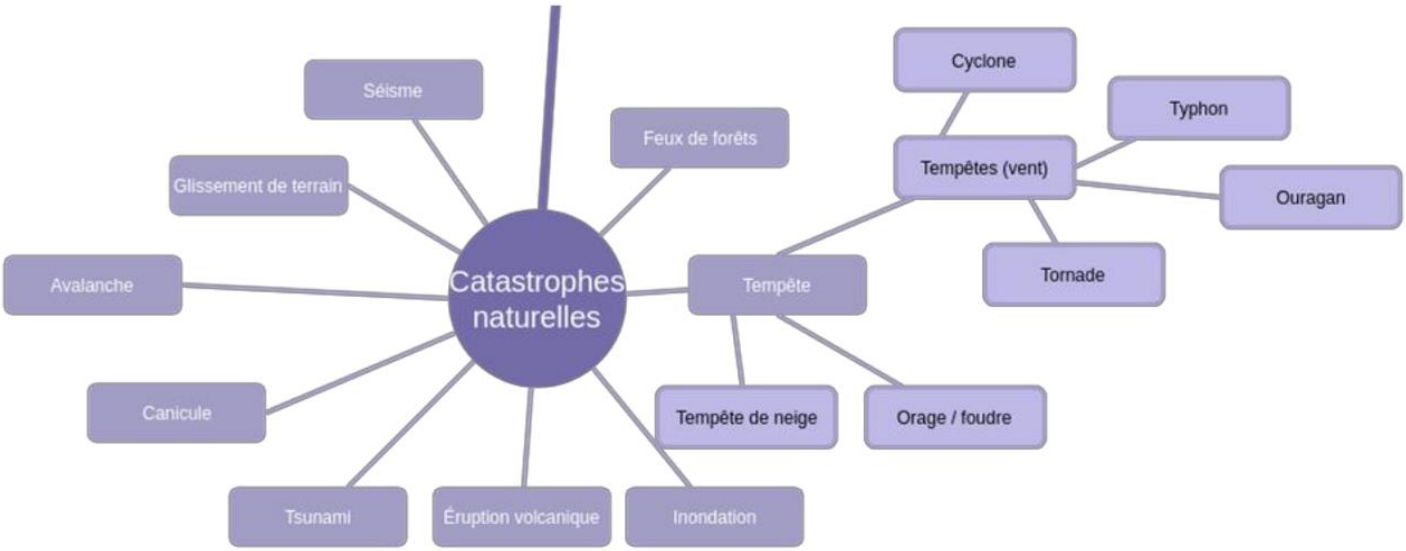

Figure 3. *Représentation graphique de la typologie d'événements liés à des décès lors de catastrophes naturelles.*

Cette typologie offre une grande diversité de représentation et de qualification des décès. Les situations sont multiples, c'est pourquoi nous nous sommes restreints, dans nos travaux ultérieurs, aux dépêches dont le type d'événement concerne les catastrophes naturelles. Cela représente un champ assez large de dépêches sans être trop volumineuses, mais aussi un type d'événement se décomposant en plusieurs sous-types. Enfin, cela nous permet de tester nos méthodes sur un périmètre réduit, ce qui permet des travaux manuels poussés, avant de l'étendre à l'ensemble des types d'événements.

Les événements utilisés par la suite sont restreints aux sous-types des catastrophes naturelles : séisme, avalanche, éruption, tsunami, canicule, inondation et tempête[6].

[6] Dans les tempêtes, nous englobons toutes les catastrophes naturelles dont l'origine est météorologique : tornade, typhon, cyclone, ouragan, orage, tempête de faible intensité, fortes pluies.

5.2. *Méthode de typage d'événements en catastrophes naturelles*

Une fois la typologie constituée, l'étape suivante consiste à pouvoir détecter les événements liés à des catastrophes naturelles. Pour ce faire, nous avons entraîné un classifieur lexical [BAR 04] à reconnaître les termes liés à chacune des catastrophes naturelles sélectionnées dans notre typologie restreinte, essentiellement par manque de données d'apprentissage. Nous avons réalisé des requêtes à un moteur de recherche afin d'enrichir le lexique [GRO 12]. Nous avons ensuite appliqué cette classification aux dépêches, ajoutant ainsi une nouvelle métadonnée aux JSON formatés, comme présenté dans la Figure 4 qui vient compléter la Figure 2.

De manière empirique, nous avons amélioré le classifieur afin de prendre en compte les mentions de « risque de » catastrophe naturelle et distinguer de manière plus précise les types de catastophes, en filtrant sur l'apparition, ou non, de certains termes. Par exemple, si dans une dépêche il est fait mention d'un séisme et d'un risque de tsunami, elle n'est catégorisée qu'en séisme, bien que des termes appartenant à la catégorie tsunami y soient présents.

```
[…]
  "category": [
    {
      "keywords": [
        "tempête",
        "tempête de neige"
      ],
      "level 1": "catastrophes_naturelles",
      "level 2": "Tempête",
      "name": "tempete",
      "score": 0.03898741550078329
    }
  ],
[…]
```

Figure 4. *Partie de la dépêche après le passage du classifieur dans laquelle nous ajoutons les types de catastrophes naturelles mentionnées dans le texte et le titre.*

5.3. *Évaluation*

Nous avons réalisé une évaluation à partir de 1 062 dépêches sélectionnées aléatoirement sur l'année 2005, pouvant concerner ou non un ou plusieurs types de catastrophes naturelles. Pour chaque dépêche, le ou les types retournés ont été évalués par un journaliste comme étant correct ou incorrect, indépendamment du nombre de types de catastrophes naturelles retournés. Le Tableau 1 présente le nombre de dépêches retournées par type de catastrophe naturelle et leurs différentes combinaisons.

Catastrophe(s) naturelle(s)	#dépêches	Catastrophe(s) naturelle(s)	#dépêches
tsunami	95	seisme tsunami inondation	3
tempete	91	seisme avalanche	2
seisme	64	risque_inondation	2
inondation	55	seisme canicule tsunami	2
avalanche	54	tsunami tempete	2
canicule	54	tsunami tempete inondation	2
inondation tempete	48	inondation seisme tsunami	1
eruption	37	risque_tsunami	1

seisme tsunami	37	risque_canicule tempete	1
seisme risque_tsunami	19	canicule inondation tempete	1
risque_avalanche	13	canicule inondation	1
avalanche inondation	8	tsunami inondation tempete	1
avalanche tempete	6	seisme tsunami tempete	1
tsunami inondation	6	tempete risque_avalanche	1
eruption inondation tempete	5	risque_inondation tsunami	1
seisme eruption tsunami	5	tempete tsunami inondation	1
risque_canicule	5	seisme inondation tempete	1
canicule tempete	4	risque_eruption	1
eruption inondation	3	seisme tempete	1
seisme eruption	3		

Tableau 1. *Nombre de dépêches par type de catastrophes naturelles.*

Au final, nous avons obtenu 93,69 % de réponses correctes. De nombreuses erreurs sont dues à la prise en compte d'un « risque » plutôt que la catastrophe naturelle en elle-même (par exemple « risque de tsunami » plutôt que « tsunami ») :

– 30 % sont dues à une inversion du « risque de catastrophe » vs. la catastrophe en elle-même ;

– 8 % sont des métaphores mal interprétées (par exemple « une avalanche d'accusations » ou « un tsunami migratoire ») ou ne concernent pas un événement (« sur une musique très romantique "la tempête de neige" ») ;

– 16 % ont un rapport étroit (par exemple « séisme » vs. « éruption ») ;

– 12 % sont partiellement correctes (c'est-à-dire qu'il manque une ou plusieurs catégories) ;

– 13 % n'ont pas donné (à tort) de catégorie ;

– 21 % sont d'autres types d'erreurs, essentiellement des extractions erronées et dues à la présence d'un lexique proche.

6. Regroupement de dépêches et identification d'événements

6.1. *Regroupement de dépêches*

L'étape suivante de nos travaux consiste à identifier les événements en question, et non plus seulement les types d'événements concernés. Pour ce faire, à partir des dépêches identifiées comme traitant de catastrophes naturelles, nous réalisons plusieurs regroupements successifs sur ce corpus réduit [KES 12] [BOS 08].

La première itération consiste à regrouper les dépêches d'une même journée en *clusters* cohérents. En effet, beaucoup de dépêches sont reprises plusieurs fois dans une même journée, voire sur deux journées consécutives, avec parfois de légères corrections ou l'ajout d'informations. Cela permet également de faire un nouveau filtrage du corpus. Pour réaliser le regroupement des dépêches, celles-ci ont été prétraitées en extrayant les noms, verbes et adjectifs du titre et du texte de la dépêche, avant de les lemmatiser puis de les représenter sous forme de sacs de mots [RIB 17]. Le regroupement a ensuite été réalisé en appliquant l'algorithme DBSCAN[7]. Le principal avantage que nous voyons à l'utilisation de DBSCAN est de ne pas avoir

[7] L'algorithme utilisé est celui implémenté dans sklearn sur python. C.f. : http://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

à définir initialement le nombre de clusters attendus. La deuxième itération voit l'application du même DBSCAN au niveau de la semaine, en sélectionnant le premier document de chaque *cluster* obtenu lors de l'itération précédente, qui représente alors, de manière arbitraire, son *cluster*. Enfin, la troisième et dernière itération applique DBSCAN au niveau du mois, de nouveau à partir de la sélection du premier document de chaque cluster obtenu lors de l'itération précédente.

6.2. *Identification d'événements*

Après avoir obtenu les regroupements finaux, nous avons cherché à identifier les événements, pour pouvoir les nommer. Cela nous permettra de rapporter l'ensemble les dépêches regroupées lors des itérations successives à un même événement, clairement identifié. Nous avons testé deux méthodes complémentaires.

La première méthode consiste à concaténer le titre le plus présent dans le cluster (un même titre pouvant être réutilisé sur plusieurs dépêches, notamment dans les comptes rendus d'événements à chaud, nécessitant des mises à jour), les dates de début et de fin du cluster et le lieu le plus mentionné dans les dépêches. Cette méthode est encore en cours d'amélioration.

La seconde méthode utilise un classifieur qui associe un *cluster* à un article Wikipédia. Cette méthode ne considère donc que les événements disposant d'un article dans l'encyclopédie en ligne, ce qui est déjà utile dans notre cas pratique et lorsque les traitements sont réalisés *a posteriori*. Les différents articles Wikipédia utilisés pour la classification des clusters ont été sélectionnés semi-automatiquement à partir des catégories de catastrophes naturelles définies dans la typologie. Chaque article sélectionné correspond à une grande catastrophe naturelle. Nous extrayons de chaque article des mots clés à partir des ancres et liens présents sur la page, ainsi que la plage temporelle sur laquelle s'est déroulée la catastrophe (c'est-à-dire les dates de début et fin). À chacune des dépêches de chaque cluster est associée un ou plusieurs articles Wikipédia en fonction des termes présents dans le texte de la dépêche et de la date. Le nom du *cluster* est alors déterminé selon le titre de l'article qui revient le plus souvent parmi les dépêches.

Les deux méthodes sont utilisées pour nommer nos *clusters*, en tentant d'abord de classer le *cluster* en fonction d'une page Wikipédia (apportant ainsi un contexte à l'événement représenté dans le *cluster*), puis, si le classifieur Wikipédia n'arrive pas à associer le *cluster* à une page, nous nommons ce dernier à l'aide de la première méthode.

6.3. *Évaluation*

Malheureusement, nous avons rapidement réalisé qu'il était complexe de définir une métrique d'évaluation unique pour cette tâche. Nous avons alors utilisé des scores de pureté (*purity scores*[8]) d'un point de vue local, ici en regroupant les clusters selon le nombre de documents qu'ils contiennent. Cela nous a amené à améliorer la qualité des résultats, mais aussi à évaluer selon les niveaux d'événement, c'est-à-dire en distinguant les événements fins et spécifiques des événements plus généraux, les seconds pouvant regrouper plusieurs des premiers événements. Afin de procéder à l'évaluation, nous avons constitué trois corpus d'évaluation composé des *clusters* obtenus sur 7 jours, 2 semaines, et deux mois. À partir de ces éléments, un journaliste a annoté chaque document de tous les *cluster*s afin de les regrouper par classe. Le Tableau 2 présente un exemple d'annotations sur le corpus hebdomadaire.

[8] https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html

Cluster Id	Titre	Texte	Classe
27	Tremblement de terre d'une magnitude de 4,8 à Santiago et dans le centre	Un séisme de magnitude 4,8 sur l'échelle de Richter a touché mardi la zone de Santiago et le centre du Chili sans faire ni victimes […]	A
27	Chili: tremblement terre magnitude 4,8 à Santiago et dans le centre pays	Un tremblement de terre de 4,8 degrés de magnitude sur l'échelle de Richter a touché mardi la zone de Santiago et […]	A
39	Nias: des survivants tirés des ruines dans le chaos des secours (REPORTAGE) Par Victor TJAHJADI =(PHOTO)=	Des survivants ont été extirpés des décombres de leur maison jeudi, trois jours après le violent séisme sur l'île indonésienne de Nias. […]	A
39	Deux avions d'aide australiens interdits d'atterrir en Indonésie	Deux avions australiens chargés d'aide se sont vu refuser jeudi l'atterrissage sur un aéroport situé dans la région indonésienne […]	B
39	Séisme : Le président indonésien au chevet des sinistrés (SYNTHESE) Par Victor TJAHJADI =(PHOTO)=	Le président indonésien Susilo Bambang Yudhoyono est arrivé jeudi sur l'île de Nias dévastée par le violent séisme de lundi […]	C
39	Séisme indonésien : une crise humanitaire "très significative" (Australie)	Le séisme survenu lundi au large de l'Indonésie représente une "crise humanitaire très significative", a déclaré jeudi […]	D
39	Séisme indonésien : plus de 620 morts, le président attendu sur place (SYNTHESE) =(PHOTO)=	Trois jours après un violent séisme qui a fait plus de 620 morts dans le nord-ouest de l'Indonésie, l'espoir s'amenuisait […]	D
46	Les deux Français disparus en Indonésie déjà dans le tsunami du 26 décembre	Les deux surfeurs français portés disparus après le séisme qui a touché l'île de Nias se trouvaient déjà dans la région de Sumatra […]	A
46	Les deux surfeurs français retrouvés sains et sauf en Indonésie	Les deux jeunes surfeurs français portés disparus après le séisme qui a frappé lundi l'île indonésienne de Nias […]	A
10	Asie/séisme: Au moins 400 morts confirmés après séisme Indonésie (responsables)	Au moins 400 personnes sont mortes après le violent séisme dans la nuit de lundi à mardi au large de l'Indonésie, dont 100 […]	A
10	Au moins 430 morts confirmés après le séisme en Indonésie (nouveau bilan)	Au moins 430 personnes sont mortes après le violent séisme dans la nuit de lundi à mardi au large de l'Indonésie […]	A
10	Nouveau choc pour les survivants traumatisés par le tsunami de décembre (PAPIER D'ANGLE) par Barry NEILD	Des millions de rescapés qui tentaient de se remettre du tsunami de décembre ont subi un nouveau choc après un violent séisme […]	B
10	De premiers secouristes aéroportés atteignent l'île de Nias	Des volontaires de la Croix-rouge indonésienne ont réussi mardi à se poser à bord d'un avion léger sur l'île de Nias […]	C
10	Indonésie: deux Suédois portés manquants sur l'île de Nias	Deux Suédois sont portés manquants après le violent séisme qui a fait au moins 430 morts dans la nuit de lundi à mardi en Indonésie […]	D
10	Indonésie: deux ressortissants de Suède portés manquants sur l'île de Nias	Deux Suédois sont portés manquants après le violent séisme qui a fait au moins 430 morts dans la nuit de lundi à mardi en Indonésie […]	D
10	En l'absence de système, le séisme a servi d'alerte au tsunami	En l'absence de système d'alerte au tsunami qui aurait été mis en place après la catastrophe de l'océan Indien de Noël […]	E

Tableau 2. *Exemple d'annotations de clusters.*

Ainsi, un cluster « pur », dont les documents sont bien, tous, censés appartenir à ce cluster, obtiendra un score de pureté de 100 % puisqu'il est composé d'une seule classe. Les annotations ont été effectuées sur chacun des corpus (jour, semaine et mois). Le Tableau 3 résume les résultats obtenus, en présentant le nombre de documents par cluster *(#docs*), le nombre de clusters concernés *(#clusters*) et le score de pureté observé.

Jours (7)			Semaines (2)			Mois (2)		
Pureté	**#docs**	**#clusters**	**Pureté**	**#docs**	**#clusters**	**Pureté**	**#docs**	**#clusters**
100 %	1	117	100 %	1	93	100 %	1	206
100 %	2	63	100 %	2	51	100 %	2	5
100 %	3	10	100 %	3	5	50 %	2	2
100 %	4	6	66,67 %	3	1	100 %	3	2
50 %	4	1	100 %	4	6	66,67 %	3	1
100 %	5	2	75 %	4	1	100 %	4	1
80 %	5	1	100 %	5	1	12,5 %	24	1
100 %	6	3	100 %	6	1			
83,33 %	6	2	42,86 %	7	1			
57,14 %	7	1	75 %	8	1			
100 %	9	2	100 %	9	3			
44,44 %	9	1	55,56 %	9	1			
100 %	10	1	100 %	10	1			
100 %	11	1	100 %	11	1			
46,15 %	13	1	54,55 %	11	2			
31,25 %	16	1	36,36 %	11	1			
54,55 %	33	1	71,43 %	14	1			
24,32 %	37	1	28,57 %	14	1			
32,14 %	56	1	50 %	18	1			
			18,42 %	38	1			
			54,35 %	46	1			
			14,29 %	56	1			
			29,51 %	61	1			

Tableau 3. *Résultats de l'évaluation sur le regroupement de dépêches.*

Le principal défaut de cette méthode est de ne fournir des résultats qu'en termes de précision, le rappel étant beaucoup trop coûteux à estimer. Il suffirait ainsi de ne fournir que de clusters composés d'un seul document pour obtenir un score de pureté de 100 %. Elle ne permet pas non plus d'avoir un score global, reflétant la qualité générale du traitement. Par contre, la méthode d'évaluation permet d'aller au détail des *clusters* évalués et ainsi améliorer de manière empirique les traitements.

Toutefois, cela nous permet par ailleurs d'estimer, par rapport à un seuil minimal de documents par cluster, le pourcentage de documents corrects (c'est-à-dire ceux se trouvant associés à des documents similaires), ainsi que le pourcentage de *clusters* corrects (c'est-à-dire ceux étant composés uniquement de documents similaires).

Le Tableau 4 présente ces résultats en fonction du nombre de documents minimum par *cluster*. La première ligne concerne ainsi l'ensemble des documents et *clusters*, la seconde sans les clusters composés d'un document, et ainsi de suite.

	Jours (7)		Semaines (2)		Mois (1)	
Seuil docs	Documents corrects	clusters corrects	Documents corrects	clusters corrects	Documents corrects	clusters corrects
1	53,16 %	94,91 %	45,87 %	91,53 %	88,46 %	98,60 %
2	52,96 %	88,89 %	45,72 %	82,14 %	36,00 %	68,75 %
3	52,56 %	69,44 %	45,4 %	54,55 %	22,81 %	57,14 %
4	51,95 %	57,69 %	44,74 %	48,15 %	8,89 %	33,33 %
5	51,12 %	47,37 %	43,71 %	35 %		
6	49,29 %	43,75 %	42,86 %	31,58 %		
7	46,76 %	36,36 %	41,8 %	27,78 %		
8			41,77 %	29,41 %		
9	46,39 %	40 %	40,91 %	31,25 %		
10	43,75 %	28,57 %	38,62 %	16,67 %		
11	40,36 %	16,67 %	36,43 %	9,09 %		
13	36,13 %	0 %				
14			32,8 %	0 %		
16	35,21 %	0 %				
18			30,6 %	0 %		
24					0 %	0 %
33	35,71 %	0 %				
37	29,03 %	0 %				
38			28,86 %	0 %		
46			31,29 %	0 %		
56	32,14 %	0 %	22,23 %	0 %		
61			29,51 %	0 %		

Tableau 4. *Résultats d'évaluation sur les clusters corrects.*

Comme l'on pourrait facilement l'imaginer, plus les clusters contiennent de document, plus la qualité s'en ressent, que ce soit en termes de documents correctement regroupés ou de clusters corrects. Toutefois, l'impact n'est pas le même. Globalement, la proportion de clusters corrects est importante, alors que la proportion de documents bien regroupés ne l'est pas, elle est même plutôt faible. Selon le cas d'usage, ce peut être l'un ou l'autre des résultats qu'il faut considérer. Dans notre cas, nous cherchons à identifier des événements, et c'est plutôt le résultat des clusters qui importe, même si l'on peut supposer que les documents mal regroupés peuvent contenir d'autres événements. Il faut également noter que les performances du regroupement au niveau de la semaine sont plus faibles que celles au niveau du jour, ce qui n'est pas étonnant du fait de l'étape supplémentaire de regroupement, qui réduit d'autant les performances.

7. Construction de patrons d'extraction et extraction de valeurs

7.1. *Principe*

Afin d'extraire des dépêches le nombre de décès présent, nous utilisons l'outil d'extraction de valeurs à partir de patrons lexico-syntaxiques décrit dans [MON 18]. Cette méthode permet, à partir de valeurs prédéfinies, de construire des patrons de phrases exprimant une relation entre ces valeurs, et ainsi d'extraire à

partir des patrons de phrases de nouvelles valeurs du même type que celles de départ. Dans un premier temps, il nous a fallu annoter un échantillon du corpus afin d'identifier les types de valeurs que nous souhaitons extraire (valeurs de départ), puis utiliser ces annotations en entraînant notre système pour construire des patrons d'extraction. Ces patrons sont ensuite testés et éventuellement modifiés, manuellement, afin d'améliorer leur capacité à retrouver les valeurs souhaitées.

7.2. *Amorçage*

Un échantillon à annoter a été constitué manuellement à partir de dépêches faisant état de décès suite à des catastrophes naturelles. Il nous a permis dans un second temps d'entraîner le système pour la construction de patrons. Nous avons tenté de trouver au moins un exemple de chacune des catégories renseignées en Section 5.1. L'échantillon regroupe au total 43 dépêches.

Les champs d'annotations correspondent à quatre des cinq « W » de la règle journalistique du même nom (*What, When, Where, Who and Why*, ce dernier étant exclu ici, son extraction étant bien trop complexe) et sont détaillés comme suit :

- *What* : Quelle catastrophe naturelle est à l'origine du (des) décès ;
- *When* : À quelle date a (ont) eu lieu le(s) décès ;
- *Where* : Où le(s) décès s'est (se sont) produit(s) ;
- *Who* : Qui est décédé (c'est-à-dire combien de personnes).

Les annotations ont été réalisées par un journaliste de WeDoData. Sur les 43 dépêches de notre corpus, le journaliste a trouvé 116 phrases candidates à l'extraction. Les informations contenues parmi ces 116 phrases sont respectivement de 63 pour le type de catastrophe, 65 pour la date, 80 pour le lieu, et enfin 112 pour le nombre de personnes décédées. Il faut toutefois noter que, parfois, les informations contenues dans les dépêches sont vagues ou imprécises (« l'an dernier », « au moins 43 personnes », « dans l'Utah », etc.). Au total, seules 32 dépêches contiennent les 4 informations à la fois, soit 28 % des 116 dépêches annotées.

7.3. *Construction semi-automatique des patrons et extraction de valeurs*

Une fois les annotations disponibles, nous avons entraîné la construction de patrons à partir des valeurs extraites des annotations. Nous avons testé les patrons obtenus sur un nouvel échantillon de test et éventuellement modifié ceux qui étaient trop précis et qui, par conséquent, ne retournaient pas assez de résultats. Des modifications ont ainsi été faites par itérations successives, ajustant les patrons à chaque passage sur le corpus.

Après avoir fait une série de tests nous avons remarqué que les patrons qui étaient limités à l'extraction du nombre de décès renvoyaient les résultats les plus variés. Nous avons donc recommencé le processus d'extraction en nous concentrant uniquement sur les extractions de nombre de morts. Au terme de la phase d'extraction de patrons nous avons finalement obtenu 19 patrons, dont la liste est donnée ci-dessous :

- *.* ont péri*
- *.* ne sont pas redescendus*
- *.* ont disparu*
- *.* ont été tuées*
- *.* a également été repêché*
- *.* a par ailleurs été tué*
- *.* sont mortes*

– *.* a également été tué*
– *.* a été retrouvé mort*
– *.* étaient mortes*
– *.* est mort*
– *Plus de .* ont été tués*
– *.* est décédé*
– *pourrait avoir fait .* morts*
– *.* sont morts*
– *.* est mort*
– *ont fait .* de morts*
– *.* morts*
– *.* victimes*

Nous appliquons nos patrons sur les dépêches et extrayons ainsi une valeur correspondant au nombre de morts. Dans un second temps, les informations contextuelles telles que le lieu et la date sont extraites au sein des phrases identifiées par le patron. Les informations de date sont extraites à l'aide du module python dateparser[9] et les informations de lieu à l'aide de l'extracteur d'entités nommées de Syllabs [MA 11], à partir de la phrase mentionnant le nombre de morts. Dans le cas où plusieurs lieux seraient présents dans la phrase, nous sélectionnons celui qui est le plus proche du nombre de morts extrait (la distance étant calculée en nombre de mots). Ainsi dans l'exemple suivant, « 40 » est extrait comme nombre de morts, « mercredi » comme date, et « Pereira » comme lieu :

« Le bilan du violent séisme qui a ravagé mercredi la région de Pereira (centre - ouest de la Colombie) a continué de s' alourdir vendredi , avec 40 morts , 281 blessés et 4.000 personnes sinistrées , selon des sources dignes de foi . »

Les informations extraites font également l'objet d'une transformation avant d'être mises en base, et ce afin d'être exploitables. Ainsi, les valeurs correspondantes au nombre de décès sont transformées en nombres entiers (par exemple « quarante » devient « 40 »), les unités extraites (« morts » dans l'exemple précédent, « personnes », « victimes », « enfants », etc.), mais également l'identification d'un facteur de certitude impliquant un nombre de morts arrondi ou imprécis (comme la présence d'expressions telles que « au moins », « près de », « plus de », etc.). Les extractions liées à la date sont analysées afin d'obtenir une date exacte par rapport à la date d'émission de la dépêche renseignée dans les métadonnées (« mercredi » devient « 1995-02-10 » dans l'exemple précédent).

7.4. *Évaluation*

Nous avons fait évaluer les résultats de 499 phrases de dépêches à un journaliste. Celui-ci devait annoter chaque phrase selon :

– l'extraction correcte ou non du nombre de décès ;

– l'extraction correcte ou non de la date ;

– l'extraction correcte ou non du lieu ;

– le typage correct ou non d'un événement ;

– la possibilité ou non de rattacher un nom de cluster à un événement, le nom étant soit un article Wikipédia, soit un regroupement d'un titre d'article ;

[9] https://pypi.org/project/dateparser/

– « l'utilisabilité » ou non de l'ensemble des informations mises à disposition pour une extraction donnée.

Ce dernier point est sans doute le plus important dans un cadre applicatif : il vise à savoir si, à partir d'un jeu de données, il est possible de finaliser un cas d'usage tel qu'une comparaison d'événement, une recherche d'informations sur un événement donné, etc.

Il a été également demandé à l'évaluateur-journaliste de considérer, en plus des valeurs binaires d'annotation, une valeur « partiel », dans le cas où l'extraction serait partiellement réalisée (par exemple le lieu « le Tyrol » plutôt que « le Tyrol et le Vorarlberg »). Les résultats sont présentés dans le Tableau 5.

	Décès	Date	Lieu	Type d'événem.	Nommage événement	Utilisabilité
#extractions	499	330	361	499	499	499
Précision [%]	94,59	83,94	80,33	80,96	53,51	52,71
Précision (partiel = correct) [%]	94,99	87,88	89,2	83,57	95,99	67,94

Tableau 5. *Résultats de l'évaluation sur l'extraction de valeurs.*

Il faut tout d'abord noter que, en ce qui concerne les extractions de date et de lieu, toutes les phrases ne sont pas complètement remplies. Ensuite, les scores de précision sont plutôt élevés et correspondent à ce que l'on pouvait attendre. La prise en compte des extractions partielles a un impact, en particulier pour les informations plus complexes à obtenir comme le nommage du *cluster*.

L'extraction des valeurs se fait en deux phases, l'extraction du groupe nominal puis l'extraction de la valeur en elle-même, et c'est cette dernière qui accentue les erreurs. Cela est principalement dû à la présence de valeurs multiples que notre extraction de valeurs, simple, a du mal à gérer, comme par exemple dans « Un homme de 58 ans » (extraction de 58), ou lorsque l'extraction de valeur ne se fait tout simplement pas. Les erreurs d'extraction sur les dates sont surtout liées à des cas particuliers (« en l'an 2.000 ») mais aussi à la non-détection de la date relative (« l'an dernier ») dû à des cas non gérés. Les erreurs sur les lieux sont plus classiques avec des extractions partielles (« Edmonds » au lieu de« Edmonds Community College ») ou de mauvais choix lors de la sélection de la valeur alors que plusieurs extractions sont présentes (« Amérique » pour « Le Salvador, le plus petit pays d'Amérique centrale).

Deux cas bien particuliers viennent expliquer les erreurs de projection sur un nom d'événement. Alors que, globalement, la projection sur une page Wikipédia se fait correctement, c'est lorsqu'il n'y a pas de page Wikipédia associée que cela se complique. Dans le cas présent, un rapprochement est réalisé avec un article du *cluster* correspondant à la dépêche en question : il semble alors évident que, dans le cas où un *cluster* est malformé ou contient des articles non liés, le risque est grand de voir apparaître des titres sans aucun rapport. Par la suite, nous prévoyons de n'utiliser que les extractions de type, de lieu et de date pour caractériser et nommer un événement. L'autre cas de figure est que nous avons utilisé les métadonnées de lieu et de date pour caractériser la dépêche. Or si la date est souvent satisfaisante (quand bien même il s'agit souvent de la date d'émission de la dépêche), le lieu est toujours celui d'où a été émis la dépêche et peut donc se situer très loin de l'action.

Enfin, l'utilisabilité est, à notre sens, la plus intéressante des métriques. Malgré sa relative réussite, elle vient nous confirmer l'intérêt de notre méthode. Certes, si nous pouvons considérer que 2/3 des extractions sont dignes d'intérêt pour l'alimentation d'une base de connaissances, la précision n'est pas optimale. Toutefois, c'est déjà une énorme avancée que de pouvoir dire à des journalistes qu'ils peuvent avoir accès à une base

structurée d'événements qu'ils peuvent traiter, et à laquelle ils peuvent se rapporter. De plus, l'utilisabilité est souvent liée au manque de données et nous avons vu précédemment les perspectives d'améliorations de nos extractions.

8. Cas d'usage : le cyclone Sidr en 2007

Nous avons regardé plus en détails les résultats des traitements pour un événement particulier : le cyclone tropical Sidr qui a traversé les districts côtiers du Bangladesh et de l'État indien du Bengale-Occidental, le 15 novembre 2007 et ayant causé la mort de plus de 3 400 personnes[10].

L'analyse des dépêches de l'année 2007 a permis de retrouver 28 extractions différentes, à partir de dépêches rédigées jusqu'à 10 jours après l'événement. 7 de ces 28 extractions étaient des doublons (qui ont bien été détectés par le système), 2 autres correspondaient à des comparaisons avec des événements passés. Les 19 extractions restantes sont présentées ci-dessous afin de montrer l'évolution du nombre de victimes lors de l'événement. Nous avons ajouté manuellement la source de l'information qui est donnée dans les dépêches.

Date et heure	#décès	Source (manuel)	Commentaire
2007-11-16 04:53:59	35	Police	La source est différente de la suivante
2007-11-16 04:53:59	100	ATN	La source est différente de la précédente
2007-11-16 05:26:00	100	ATN	La source est différente de la suivante
2007-11-16 05:26:00	150	Armée	La source est différente de la précédente
2007-11-16 11:00:12	200	Des responsables	
2007-11-16 13:25:06	550	Télévision locale	
2007-11-17 12:58:26	1070	Responsable officiel	
2007-11-17 14:49:55	1723	Armée	
2007-11-17 17:37:36	1700	*Inconnue*	
2007-11-18 11:33:14	2000	Gouvernement bangladais	
2007-11-18 12:33:52	2217	Gouvernement bangladais	
2007-11-18 13:00:54	2200	*Inconnue*	
2007-11-18 14:23:36	2217	Gouvernement bangladais	
2007-11-18 17:43:39	2300	Gouvernement bangladais	La source est différente de la suivante
2007-11-18 17:43:39	3000	Croissant - Rouge bangladais	La source est différente de la précédente
2007-11-19 00:00:01	10000	*Inconnue*	La valeur correcte est « entre 5 et 10000 »
2007-11-20 00:08:40	10000	Croissant - Rouge bangladais	La valeur correcte est « entre 5 et 10000 »
2007-11-20 13:07:14	10000	Croissant - Rouge bangladais	La valeur correcte est « entre 5 et 10000 »
2007-11-26 07:25:58	3447	« Officiellement »	

Tableau 6. *Analyse des dépêches liées au cyclone Sidr.*

[10] https://fr.wikipedia.org/wiki/Cyclone_Sidr

Dans un cas d'usage où un humain aurait à utiliser et analyser ces données, nous pouvons voir qu'il serait possible de montrer l'évolution de la situation et même d'en faire l'analyse, ne serait-ce qu'en comparant les sources. Nous pouvons également imaginer une comparaison avec d'autres événements similaires. La figure suivante montre l'évolution du nombre de décès de manière graphique.

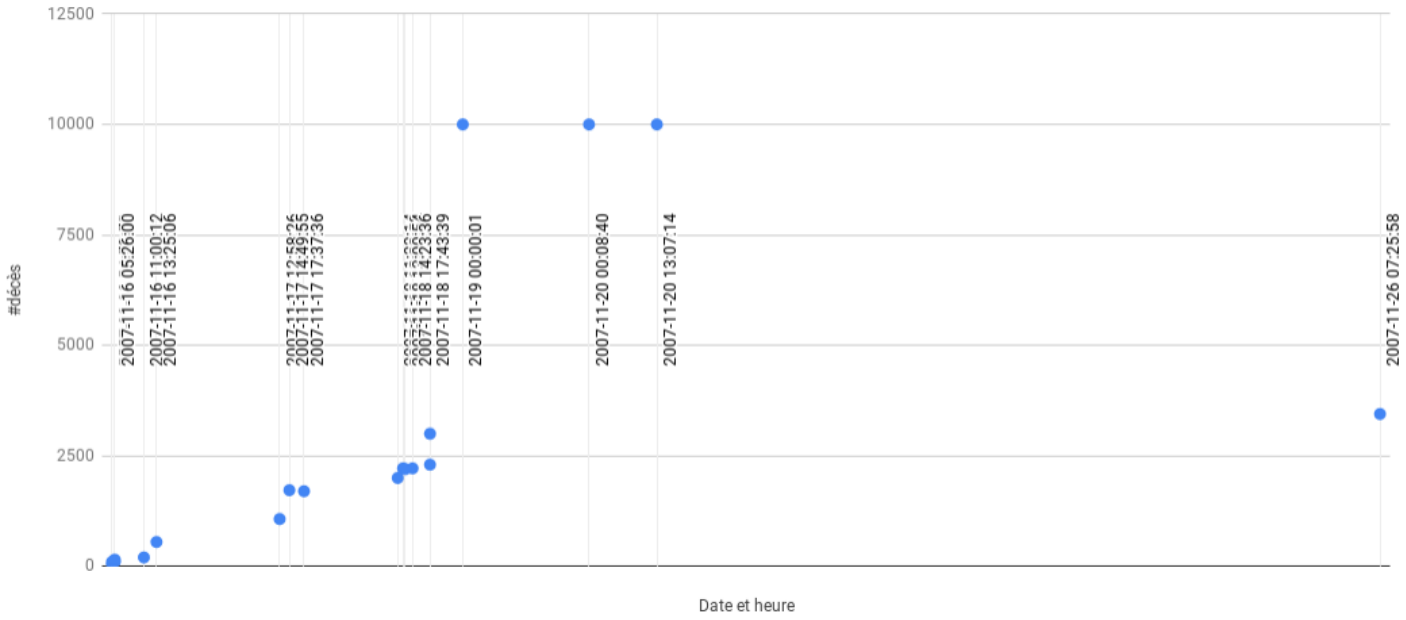

Figure 5. *Évolution du nombre de décès lors du cyclone Sidr dans les dépêches AFP.*

L'analyse des quelques dépêches liées au cyclone Sidr nous permet d'une part de visualiser l'évolution du nombre de décès au cours du temps, mais également selon les sources. L'accès à un tel jeu de valeurs est important, quitte à imaginer un post-traitement manuel suivant les extractions. Le cas du Croissant Rouge Bangladais est particulièrement enrichissant car il vient rompre les estimations officielles. L'évolution au cours du temps et l'espacement dans le temps des dépêches est également intéressant : le décompte augmente régulièrement sur deux jours, puis un bilan radicalement différent est donné d'une source différente, et enfin, une semaine plus tard, un bilan final est donné (depuis la première source).

En revanche, le fait d'avoir plusieurs chiffres est problématique si l'on veut travailler sur des comparaisons entre catastrophes naturelles. Dans ce cas, il faut disposer du dernier bilan officiel. Une validation manuelle sera alors nécessaire et même un système automatique peut aider dans cette tâche.

9. Conclusion

Dans cet article, nous avons présenté une méthodologie pour l'extraction de valeurs dans un corpus de dépêches, en vue d'utiliser ces valeurs dans un contexte journalistique. Le domaine restreint concerne les décès lors de catastrophes naturelles.

Dans un premier temps, nous avons cherché à détecter les types de catastrophes naturelles. La méthode utilisée a montré de bons résultats, à hauteur de 94 % de réussite.

Puis, nous avons réalisé des regroupements successifs par jour, semaine et mois, et ce afin réduire le volume de dépêches à traiter, regrouper ces dépêches sous forme d'événements, mais aussi parce que ces événements peuvent courir sur plusieurs jours. Les résultats obtenus sont globalement positifs, même s'ils montrent la disparité des regroupements réalisés.

Au-delà du fait de regrouper les dépêches par événement, un des challenges que nous avons entrepris fût de nommer ces événements. Cela s'est fait en deux étapes : d'une part un lien vers un article Wikipédia, en particulier pour des événements connus, puis en utilisant les titres pour illustrer au maximum les dépêches. Si le lien vers Wikipédia est plutôt utile, les résultats nous ont montré que notre méthode de nommage à partir des titres est à revoir.

Enfin, nous avons utilisé des patrons lexico-syntaxique en vue d'extraire des valeurs en lien avec des décès de personnes. En conservant les phrases dans lesquelles apparaissent ces valeurs, nous en avons également extrait un contexte : la date, le lieu, la nuance ainsi que l'unité. Ces données sont utiles en soit et apportent une vraie valeur ajoutée pour la constitution d'une base de connaissances. L'évaluation réalisée en termes d'utilisabilité est particulièrement parlante et nous sommes très optimistes quant à l'intérêt de nos travaux.

Dans le cas précis du corpus AFP que nous avons utilisé, 38 millions de dépêches ont été filtrées en 11 millions de dépêches en français, puis traitées pour obtenir 151 370 événements de type catastrophe naturelle (pour 52 375 événements distincts), et enfin 66 885 extractions de valeurs parmi 40 708 dépêches. L'ensemble a été déposé dans une base de données pour une utilisation ultérieure.

Dans le futur, nous souhaitons appliquer notre méthode sur les autres types de catastrophes, puis de décès. Ensuite, nous l'ouvrirons à d'autres domaines fournissant des types de valeurs différentes, comme des prix, des volumes, hauteurs, surfaces, etc.

Les valeurs extraites servent à remplir une base de connaissances pouvant être utilisée par des journalistes, par exemple dans un contexte de comparaison de valeurs (par exemple en retrouvant le bilan d'un séisme similaire à un séisme qui vient de se produire). D'autres applications sont susceptibles de voir le jour, comme un explorateur d'événements, un explorateur de dépêches, ou encore un validateur (factuel) de dépêches. Ces applications, pilotées par la donnée, pourront enrichir le message qui est donné par les journalistes.

Références

[AKB 12] AKBIK A., VISENGERIYEVA L., HERGER P., HEMSEN H., LÖSER A., « Unsupervised Discovery of Relations and Discriminative Extraction Patterns », *COLING*, 2012.

[BAR 04] BARONI M., BERNARDINI S., « BootCaT: Bootstrapping corpora and terms from the web », *LREC*, 2004.

[BOS 08] BOSSARD A., POIBEAU T., « Regroupement Automatique de Documents en Classes Événementielles », *TALN*, 2008.

[BRI 98] BRIN S., « Extracting patterns and relations from the World Wide Web », *International Workshop on the World Wide Web and Databases*, 1998.

[CAF 05] CAFARELLA M., DOWNEY D., SODERLAND S., ETZIONI O., « KnowItNow: Fast, scalable information extraction from the web », *Human Language Technology and Empirical Methods in Natural Language Processing*, 2005.

[GRO 12] DE GROC C., TANNIER X., DE LOUPY C., « Un critère de cohésion thématique fondé sur un graphe de cooccurrences », *JEP-TALN-RECITAL*, 2012.

[ETZ 04] ETZIONI O., CAFARELLA M., DOWNEY D., KOK S., POPESCU A., SHAKED T., SODERLAND S., WELD D., YATES A., « Web-scale information extraction in knowitall », *13th international conference on World Wide Web*, 2004.

[ETZ 11] ETZIONI O., FADER A., CHRISTENSEN J., SODERLAND S., MAUSAM M., « Open Information Extraction: The Second Generation », *IJCAI*, 2011.

[GRI 96] GRISHMAN R., SUNDHEIM B., « Message Understanding Conference - 6: A Brief History », *COLING*, 1996.

[LIN 01] LIN D., PANTEL P., « DIRT discovery of inference rules from text », *ACM SIGKDD*, 2001.

[MA 11] MA J., MOUNIER M., BLANCAFORT H., COUTO J., DE LOUPY C., « LOL: Langage objet dédié à la programmation linguistique », *TALN*, 2011.

[KES 12] KESSLER R., TANNIER X., HAGEGE C., MORICEAU V., BITTAR A., « Extraction de dates saillantes pour la construction de chronologies thématiques », *TAL*, Volume 52 – n°2/2012, 2012.

[MON 18] MONNIN C., HAMON O., « Construction de patrons lexico-syntaxiques d'extraction pour l'acquisition de connaissances à partir du web », *CORIA-TALN-RJC*, 2018.

[MOR 99] MORIN E., « Extraction de liens sémantiques entre termes à partir de corpus de textes techniques », *Thèse de doctorat. Nantes*, 1999.

[RIB 17] RIBEIRO S., FERRET O., TANNIER X., « Unsupervised Event Clustering and Aggregation from Newswire and Web Articles », *NLPmJ@EMNLP*, 2017.