# Construire et évaluer une application de veille pour l'information sur les événements sismiques

**Romaric Besançon — Olivier Ferret — Ludovic Jean-Louis**

*CEA, LIST, Laboratoire Vision et Ingénierie des Contenus*
*Fontenay-aux-Roses, F-92265, France.*
*{romaric.besancon,olivier.ferret,ludovic.jean-louis}@cea.fr*

RÉSUMÉ. *Le développement d'applications opérationnelles de veille pour des domaines spécifiques nécessite l'intégration de nombreuses techniques et outils issus du champ de la recherche d'information et du traitement automatique des langues. Dès lors, un des défis présidant à une telle intégration est la prise en compte des limitations propres à chacune de ces techniques et outils en termes d'influence sur le résultat final du système. Plus précisément, nous présentons dans cet article une application pour la surveillance des informations concernant les événements sismiques sur le Web. Nous nous attachons à présenter les différents composants nécessaires à une telle application, les solutions proposées pour ces différents composants et une évaluation indépendante de chacune de ces solutions ainsi qu'une analyse de leur influence sur le résultat global du système.*

ABSTRACT. *Operational intelligence applications in specific domains are developed using numerous natural language processing technologies and tools. A challenge for this integration is to take into account the limitations of each of these technologies in the global evaluation of the application. We present in this article an intelligence application for the gathering of information from the Web about recent seismic events. We present the different components needed for the development of such system and the technologies behind each component. We also propose an independent evaluation of each component in order to analyze their influence in the overall performance of the system.*

MOTS-CLÉS : *Application industrielle, extraction d'information, veille en source ouverte, reconnaissance d'entités nommées*

KEYWORDS: *Industrial application, information Extraction, open-source intelligence, named entitity recognition*

## 1. Introduction

L'application présentée dans cet article est issue d'un projet de veille événementielle pour les besoins des analystes du Département d'Analyse et Surveillance de l'Environnement au CEA, qui a en particulier une mission de veille sur les événements sismiques. L'objectif de l'application de veille proposée est d'assister ces analystes pour relier les informations sismiques repérées sur des sismographes à des informations publiées dans des dépêches de presse sur Internet. Cette assistance a pour objet de repérer des entités spécifiques dans les textes (lieux, dates, magnitudes), et d'exploiter ces entités pour retrouver les événements et les présenter à l'utilisateur.

Dans le domaine de la surveillance des événements sismiques, certaines études ont montré l'intérêt d'enrichir la détection directe par un repérage des événements dans des textes (Sakaki *et al.,* 2010, Earle *et al.,* 2010). Ces études utilisent plus particulièrement des flux d'informations courtes (Twitter), en utilisant leur aspect temporel et éventuellement spatial : étant donnée la taille des *tweets*, la détection d'informations dans le texte est sommaire ; c'est la masse qui fournit l'information. L'application que nous proposons ici s'appuie plutôt sur une analyse profonde de textes plus structurés (dépêches d'actualité) et sont moins adaptés à la détection d'événements sismiques en temps réel qu'à une aide à l'interprétation des événements détectés sur les capteurs (confirmation d'une détection, informations de ressenti, dégâts causés).

Dans ce contexte, notre application représente un exemple typique d'application d'extraction d'information (Cunningham, 2005) en domaine spécialisé, dont le but est la détection d'une structure associant plusieurs entités en relation dans la description d'un événement (*scenario template*, pour reprendre la terminologie de MUC (Grishman *et al.,* 1996)). Ce type d'applications nécessite l'intégration de différents niveaux de traitement automatique de la langue, incluant la collecte de textes informatifs (extraction de contenu), le filtrage d'information, la reconnaissance des entités nommées et l'identification des relations entre ces entités. Nous présentons dans cet article les solutions utilisées pour ces différents traitements, en proposant des évaluations individualisées de chacun d'eux afin de mettre en évidence leurs limitations spécifiques et leur influence sur les performances globales de l'application.

## 2. Aperçu général de l'application de veille

L'objectif de l'application est de repérer, dans des textes publiés sur le Web, l'évocation d'un événement sismique et d'identifier de façon automatique les informations qui y sont associées. Nous présentons ici les différentes étapes nécessaires pour atteindre cet objectif, qui sont détaillées dans les sections suivantes :

– les dépêches sont collectées à partir des sources d'information choisies, dans leur format brut, puis le contenu textuel des dépêches est extrait à partir de ce format brut (section 3) ;

– l'événement est repéré dans le texte de la dépêche suivant un processus en trois étapes : une détection des entités spécifiques du domaine (section 4), une segmentation du texte en événements et finalement, le rattachement des entités pertinentes à l'événement principal (section 5) ;

– les dépêches non pertinentes sont filtrées selon deux critères : le premier est la détection effective d'un événement dans le texte par le module précédent ; le second utilise un outil de classification statistique (section 6) ;

## 3. Collecte et nettoyage des textes

Dans le domaine de cette application spécifique de veille, nous nous intéressons à la collecte d'informations d'actualité provenant d'articles de journaux ou de dépêches d'agences de presse (les sources d'information rapide de type Twitter n'ont pas été exploitées). Ces dépêches peuvent être obtenues directement par l'abonnement à un fil d'actualité auprès d'une agence de presse ou récupérées sur le Web. Dans le premier cas, les dépêches sont dans un format structuré – par exemple, l'Agence France Presse (AFP) fournit des dépêches au format NewsML– et les différentes informations utiles sont donc accessibles facilement. Dans le second cas, les dépêches sont récupérées sous forme de pages Web, dont le format n'est pas connu, et nécessitent donc un traitement spécifique pour en extraire la partie informative (le texte de la dépêche) et filtrer le reste (titres, menus, publicités, etc.).

Après une première étape de nettoyage de l'encodage des pages (détection de l'encodage de la page et ré-encodage éventuel en UTF8), deux méthodes ont été testées pour extraire le contenu textuel informatif de la page HTML, nommées *density-cleaner* et *html-cleaner*. *Density-cleaner* convertit tout d'abord le HTML en texte (en utilisant le navigateur texte Lynx), puis détecte la partie informative en s'appuyant sur l'idée que les parties non informatives contiennent plus de structures (listes, tableaux) et sont donc moins denses en texte. Elle recherche donc les plus fortes ruptures de densité de texte, en utilisant une fenêtre glissante à la manière de (Hearst, 1997) pour la segmentation thématique : la première plus forte augmentation de densité trouvée indique le début de la zone informative, la plus forte diminution de densité suivante indique la fin de la zone. *Html-cleaner* détecte quant à elle les zones de texte directement dans le code HTML, en cherchant des indices de présence de texte (tels que les balises <*p*>, la présence de ponctuations, etc.) et on utilise la structure HTML du document pour remonter à la balise englobante du texte. L'utilisation de la balise englobante permet d'éviter le problème de la détection d'une frontière droite mal définie auquel on peut être confronté avec la première méthode. Les indices pour la détection de zones de textes ont été inspirés par l'algorithme sous-tendant l'outil Web *Readability*[1].

Pour l'évaluation du prétraitement des dépêches, nous avons utilisé le format et les outils d'évaluation de la campagne d'évaluation CLEANEVAL sur le nettoyage de pages Web (Baroni *et al.,* 2008), sur un corpus de référence construit en annotant 50

1. Readability - An Arc90 Lab Experiment (http://lab.arc90.com/experiments/readability/)

pages Web collectées par notre application (correspondant à 5600 lignes de textes). Le tableau 1 présente les résultats de l'évaluation des deux méthodes testées sur ce corpus et donne en comparaison les résultats de l'outil de nettoyage *Ncleaner* (Evert, 2008) (utilisé avec le modèle par défaut, *i.e.* sans apprentissage spécifique lié à notre corpus, qui devrait aboutir à de meilleurs résultats).

| | Précision | Rappel | F-mesure |
|---|---|---|---|
| density-cleaner | 56,5% | 90,4% | 69,5% |
| html-cleaner | 96,5% | 94,4% | 95.4% |
| Ncleaner | 64,9% | 83,7% | 73,1% |

**Tableau 1.** *Évaluation de l'extraction du contenu informatif des pages Web*

La méthode de nettoyage utilisant la structure HTML des pages et des indices de présence de textes donne de bien meilleurs résultats globaux. Il est toutefois à noter qu'avec cette méthode, les résultats sont moins réguliers : d'une part, lorsque le repérage du contenu échoue, on perd souvent toute l'information (on prend une tout autre partie du document HTML) ; d'autre part, pour certaines dépêches, on garde toute la page (le noeud HTML gardé est le noeud racine). Malgré ces défauts, les résultats restent meilleurs et sur un corpus de 500 dépêches, nous avons estimé à 1% des dépêches les occurrences de chacun de ces cas d'erreur.

## 4. Reconnaissance des entités spécifiques

Les dépêches sont analysées en utilisant l'analyseur linguistique LIMA du CEA LIST (Besançon *et al.,* 2010). Le module de reconnaissance des entités spécifiques intégré dans LIMA fonctionne sur la base de règles développées manuellement s'appuyant sur des listes de déclencheurs et sur le contexte local autour de ces annonceurs ainsi que sur des listes d'entités existantes (*e.g.* noms de lieux) construites semi-automatiquement. Les entités spécifiques d'intérêt pour notre application ont été définies avec des analystes du domaine et sont : le type d'événement (EVENT_TYPE), le lieu de l'événement (LOCATION), la date et l'heure de l'événement (DATE/TIME), sa magnitude (MAGNITUDE), les dégâts causés (DAMAGES) et les coordonnées géographiques de l'événement (GEO_COORD).

Pour l'évaluation de la reconnaissance des entités nommées, deux corpus ont été utilisés : un corpus de 50 dépêches annotées entièrement et un corpus de 501 dépêches avec une annotation partielle des entités dans laquelle seules les entités liées à l'événement principal ont été identifiées[2]. Pour cette seconde évaluation, seul le rappel est mesuré (les entités n'étant pas toutes annotées dans la référence, la précision n'a pas de pertinence). Les résultats des deux évaluations sont présentés dans le tableau 2.

2. Cette annotation de référence a été faite par les analystes du Laboratoire de détection géophysique du CEA.

| | complet_50 | | | partiel_501 |
|---|---|---|---|---|
| type d'entité | Précision | Rappel | F-mesure | Rappel |
| EVENT_TYPE | 93,9% | 93,0% | 93,4% | 97,4% |
| LOCATION | 90,5% | 66,5% | 76,6% | 84,4% |
| DATE | 88,2% | 86,3% | 87,2% | 98,7% |
| TIME | 82,6% | 86,5% | 84,5% | 96,5% |
| MAGNITUDE | 93,8% | 83,3% | 88,2% | 94,0% |
| DAMAGES | 83,5% | 63,9% | 72,4% | 62,7% |
| GEO_COORD | 100,0% | 66,7% | 80,0% | 86,7% |
| toutes entités | 89,8% | 77,4% | 83,2% | 72,9% |

**Tableau 2.** *Évaluation de la reconnaissance des entités spécifiques sur 50 dépêches annotées complètement et sur 501 dépêches annotées partiellement.*

Mêmes si les résultats sont plutôt inférieurs aux résultats de l'état de l'art pour la reconnaissance des entités nommées, les performances sont acceptables. En particulier, le taux de rappel sur le corpus de 501 dépêches est très bon, ce qui est encourageant pour les traitements postérieurs s'appuyant sur ces résultats. Les moins bons scores sur une entité comme DAMAGES s'expliquent par la plus grande variabilité d'expression pour cette entité (pour laquelle les annotations de référence comptent des expressions simples comme « *65 blessés* » et des expressions plus complexes comme « *faisant six morts, interrompant la circulation des trains, provoquant des glissements de terrain et l'effondrement d'un pont* »).

## 5. Identification des événements et rattachement des entités

Nous nous intéressons dans cette application à l'extraction d'événements ponctuels et datés. Or, les textes considérés, de par leur nature journalistique, font souvent référence à plusieurs événements de même nature en comparant l'impact d'un événement sismique récent à celui d'un tremblement de terre important passé. Nous proposons donc, comme premier temps de notre processus d'extraction d'information, de structurer le texte en segments temporellement homogènes dans lesquels un seul événement est évoqué. Dans un second temps, à l'intérieur de la section la plus intéressante (celle relative à un événement et ayant la date la plus récente), nous recherchons les entités les plus pertinentes et nous les rattachons à la structure de l'événement.

Pour l'étape de segmentation, deux méthodes ont été testées : (1) une méthode heuristique exploitant la présence et la valeur des dates selon deux principes : des dates ayant des valeurs différentes correspondent à des segments différents (le segment principal étant celui de date la plus récente) ; les ruptures de segments entre deux dates différentes s'appuient sur la structure du texte en phrases et paragraphes ainsi que sur la présence d'autres entités caractéristiques du domaine entre les dates (2) une méthode à base d'apprentissage automatique classant les phrases du texte en trois

catégories (« *événement principal* », « *événement secondaire* » et « *contexte* ») en se fondant sur la séquence des temps des verbes, la présence des dates et la présence d'autres expressions temporelles (Jean-louis *et al.,* 2010). Le modèle d'apprentissage donnant les meilleurs résultats pour cette tâche est celui des Champs Conditionnels Aléatoires (CRF). Pour l'étape de rattachement, nous utilisons pour le moment une heuristique simple : pour chaque type d'entité, nous prenons la première entité dans la section de l'événement principal. Cette heuristique repose sur l'idée – souvent vérifiée en pratique – que dans une dépêche d'actualité, l'événement principal est évoqué avant un événement secondaire.

Une évaluation intrinsèque de la segmentation en événements a été effectuée sur un corpus de 140 dépêches, sélectionnées principalement parmi celles évoquant plusieurs événements et annotées en segments événementiels (*i.e.* en segments homogènes correspondant à un même événement). Les résultats de cette évaluation sont présentés

| | Heuristique | | CRF | |
|---|---|---|---|---|
| Type d'événement | Rappel | Précision | Rappel | Précision |
| Événement principal | 82,8% | 64,7% | 98,7% | 87,4% |
| Événement secondaire | 23,5% | 43,4% | 52,7% | 95,8% |
| Contexte | 16,9% | 21,7% | 69,3% | 92,7% |

**Tableau 3.** *Résultats de la segmentation en événements*

dans le tableau 3. Les résultats de l'évaluation de l'impact de la segmentation sur le rattachement des entités aux événements sont présentés dans le tableau 4, pour les deux stratégies de segmentation, en les comparant avec les résultats sans segmentation et en utilisant la segmentation de référence.

| | Sans segmentation | | Heuristique | | CRF | | Segm. référence | |
|---|---|---|---|---|---|---|---|---|
| Type d'entité | Rappel | Préc. | Rappel | Préc. | Rappel | Préc. | Rappel | Préc. |
| EVENT_TYPE | 82,1% | 81,6% | 79,3% | 78,8% | 76,7% | 76,2% | 85.6% | 85.6% |
| LOCATION | 41,0% | 40,9% | 56,0% | 55,9% | 57,4% | 57,3% | 86.4% | 86.4% |
| DATE | 38,4% | 35,9% | 69,3% | 65,0% | 64,4% | 60,1% | 89.5% | 86.9% |
| TIME | 61,1% | 51,2% | 56,4% | 49,2% | 63,4% | 55,5% | 92.2% | 91.3% |
| MAGNITUDE | 93,5% | 93,0% | 86,3% | 85,9% | 86,7% | 86,1% | 93.4% | 93.4% |
| DAMAGES | 83,5% | 77,9% | 76,3% | 74,4% | 80,2% | 75,3% | 76.7% | 73.5% |
| GEO_COORD | 86,7% | 96,3% | 66,7% | 74,1% | 83,3% | 92,6% | 100% | 100% |
| Tous | 66,6% | 63,5% | 71,0% | 68,6% | 71,7% | 68,8% | 87.5% | 86.3% |

**Tableau 4.** *Impact de la segmentation sur le rattachement des entités à l'événement principal*

Ces résultats montrent que l'heuristique de rattachement, même simple, donne déjà des résultats satisfaisants, mais que l'étape de segmentation des textes apporte néanmoins une réelle amélioration. La différence entre les deux techniques de segmentation testées n'est pas significative sur ces résultats globaux et leur différence

par rapport au résultat obtenu avec la segmentation de référence montre que les résultats de segmentation doivent encore pouvoir être améliorés. Une analyse plus fine des résultats nous a montré qu'avec des pourcentages d'entités correctement rattachées proches entre ces deux cas, les erreurs de segmentation sont en fait divisées par deux avec la méthode de segmentation par CRF, les erreurs étant alors reportées sur des erreurs de rattachement à l'intérieur du segment principal. Ces erreurs restantes doivent être traitées en envisageant des stratégies de rattachement plus sophistiquées, utilisant par exemple les proximités entre entités ou les structures syntaxiques qui les relient.

## 6. Filtrage des dépêches

Le filtrage des dépêches permet de ne conserver, dans l'affichage synthétique des informations, que des dépêches pertinentes pour les analystes du domaine. Le premier critère de filtrage est la détection effective d'un événement par le module d'identification des événements. Mais en pratique, cette détection n'est pas suffisante : nous avons mesuré qu'il reste après ce premier filtrage 60% de dépêches non pertinentes. Parmi ces dépêches, certaines utilisent par exemple les mots du domaine (« *tremblement de terre* ») dans un sens imagé ; d'autres évoquent un événement réel, mais de façon anecdotique. Un second filtrage a donc été intégré, utilisant un système de classification statistique entraîné sur un corpus annoté manuellement. Nous avons classiquement utilisé pour ce faire un classifieur de type *Machines à Vecteurs de Support* (*SVM*) avec un corpus d'entraînement constitué de 501 dépêches pertinentes et 711 dépêches non pertinentes et un corpus de test distinct de 91 dépêches pertinentes et 166 dépêches non pertinentes.

Une étude de ses paramètres a été effectuée pour optimiser ses performances : en particulier, l'utilisation de mots simples ou de lemmes comme unités de représentation dans la similarité des textes et la pondération de ces unités, binaire (présence/absence) ou fréquentielle (*tf.idf*). Une optimisation du seuil de décision pour les SVM a également été utilisée, le seuil par défaut du modèle (=0) n'étant pas forcément le meilleur (Shanahan *et al.,* 2003). Le tableau de résultats 5 montre que le meilleur compromis entre précision et rappel est obtenu en utilisant les lemmes, avec une pondération binaire. L'option *lemmes-tf.idf* a une meilleure *accuracy*, mais le rappel est plus bas et le filtrage est alors jugé trop strict par les utilisateurs.

| | Précision | Rappel | F-mesure | Accuracy |
|---|---|---|---|---|
| mots / présence | 80,6% | 91,2% | 85,6% | 89,1% |
| mots / tf.idf | 93,5% | 79,1% | 85,7% | 90,7% |
| lemmes / présence | 84,5% | 91,1% | 87,7% | 91,0% |
| lemmes / tf.idf | 93,6% | 81,1% | 86,9% | 91,4% |

**Tableau 5.** *Évaluation des différentes stratégies de filtrage pour des SVM, avec optimisation du seuil de décision*

## 7. Conclusion

Cet article présente une application industrielle de veille événementielle dans le domaine des événements sismiques. Pour être opérationnelle, cette application intègre plusieurs composants de traitement et d'analyse des textes : des composants amont de filtrage et nettoyage de textes et des composants effectifs d'extraction d'information prenant en compte la spécificité des documents du domaine. En particulier, le fait, pour les dépêches émises, d'évoquer souvent plusieurs événements ajoute de l'ambiguïté dans l'identification des entités liées à l'événement principal. L'évaluation globale d'une application intégrant autant de composants est délicate car chacun des composants possède ses propres contraintes et limitations. Nous avons proposé dans cet article une évaluation quantitative individualisée de chacun des composants pour mettre en évidence les limitations et les pertes qui leur sont dues dans l'évaluation globale de l'application, telle qu'elle peut être ressentie par un utilisateur. Ces évaluations permettent d'améliorer les différents composants en mettant en évidence les points faibles de l'application. Dans cette optique, nos prochains efforts porteront sur l'identification des événements par le développement de stratégies plus complexes pour le rattachement des entités aux événements se fondant sur des critères de proximité et des critères linguistiques (relations syntaxiques entre les entités).

## 8. Bibliographie

Baroni M., Chantree F., Kilgarriff A., Sharoff S., « Cleaneval : a Competition for Cleaning Web Pages », *Proceedings of LREC'08*, 2008.

Besançon R., de Chalendar G., Ferret O., Gara F., Mesnard O., Laïb M., Semmar N., « LIMA : A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation », *Proceedings of LREC'10*, 2010.

Cunningham H., *Encyclopedia of Language and Linguistics*, Elsevier, chapter Information extraction, automatic, 2005.

Earle P., Guy M., Buckmaster R., Ostrum C., Horvath S., Vaughan A., « OMG Earthquake ! Can Twitter Improve Earthquake Response ? », *Seismological Research Letters*, vol. 81, n° 2, p. 246-251, 2010.

Evert S., « A Lightweight and Efficient Tool for Cleaning Web Pages », *Proceedings of LREC'08*, 2008.

Grishman R., Sundheim B., « Message Understanding Conference-6 : a brief history », *Proceedings of the 16th conference on Computational linguistics - Volume 1*, p. 466-471, 1996.

Hearst M. A., « TextTiling : Segmenting text into multi-paragraph subtopic passages », *Computational Linguistics*, vol. 23, n° 1, p. 33-64, 1997.

Jean-louis L., Besançon R., Ferret O., « Utilisation d'indices temporels pour la segmentation événementielle de textes », *Actes de la conférence TALN 2010*, 2010.

Sakaki T., Okazaki M., Matsuo Y., « Earthquake shakes Twitter users : real-time event detection by social sensors », *Proceedings of WWW'10*, ACM, p. 851-860, 2010.

Shanahan J. G., Roma N., « Improving SVM Text Classification Performance through Threshold Adjustment », *Proceedings of ECML'2003*, p. 361-372, 2003.