HAL open science

Détection d'informations vitales pour la mise à jour de bases de connaissances

Rafik Abbes, Nathalie Jane Hernandez, Karen Pinel-Sauvagnat, Mohand Boughanem

To cite this version:

Rafik Abbes, Nathalie Jane Hernandez, Karen Pinel-Sauvagnat, Mohand Boughanem. Détection d'informations vitales pour la mise à jour de bases de connaissances. Conférence d'Ingénierie des Connaissances (IC 2015), Jun 2015, Rennes, France. hal-01165507v1

HAL Id: hal-01165507

https://inria.hal.science/hal-01165507v1

Submitted on 19 Jun 2015 (v1), last revised 24 Jun 2015 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Détection d'informations vitales pour la mise à jour de bases de connaissances

Rafik Abbes, Nathalie Hernandez, Karen Pinel-Sauvagnat, Mohand Boughanem

INSTITUT DE RECHERCHE EN INFORMATIQUE DE TOULOUSE
118, ROUTE DE NARBONNE, F-31062 TOULOUSE CEDEX 9, FRANCE
RAFIK.ABBES, NATHALIE.HERNANDEZ, KAREN.SAUVAGNAT, MOHAND.BOUGHANEM @IRIT.FR

Résumé : Mettre à jour une base de connaissances est une problématique actuelle qui suit l'évolution permanente du web de données liées. De nombreuses approches ont été proposées afin d'extraire dans des documents textuels la connaissance à mettre à jour. Ces approches arrivent à maturité mais reposent sur l'hypothèse selon laquelle le corpus adéquat a déjà été constitué. Dans la majorité des cas, les documents à prendre en compte sont sélectionnés manuellement ce qui rend difficile une mise à jour exhaustive de la base. Dans cet article nous proposons une approche originale visant à identifier automatiquement dans un flux de documents du web les éléments pouvant apporter de la connaissance nouvelle sur des instances déjà représentées dans une base.
Mots-clés : Informations vitales, Mise à jour de bases de connaissances, TREC Temporal Summarization

1 Introduction

Les bases de connaissances telles que *DBpedia* sont devenues des sources indispensables pour rendre accessibles à tout un chacun la connaissance relative aux instances du monde réel comme les personnes, les organisations, les événements, etc. Au cours du temps, la connaissance relative à ces instances peut évoluer lorsque, par exemple, dans le cas de personnes, ces instances réalisent de nouvelles actions, ou se trouvent dans de nouvelles situations. Ceci implique un travail permanent de suivi pour maintenir les bases de connaissances à jour.

L'extraction de connaissances à partir de documents textuels est une approche couramment utilisée pour la constitution de base de connaissances [Petasis *et al.* (2011)]. Ces approches reposent souvent sur l'hypothèse selon laquelle le corpus à partir duquel la connaissance est extraite est identifié, que ce soit à partir des pages Wikipedia dans le cas de DBpedia, ou constitué manuellement [Augenstein *et al.* (2012); Exner & Nugues (2012)]. Dans le contexte de la mise à jour de bases de connaissances, la tâche d'identification des textes d'où extraire la connaissance n'est pas triviale. D'une part, certaines approches d'extraction de connaissance à partir du texte analysent les documents dans leur intégralité, or la connaissance sur une instance donnée est souvent décrite uniquement dans quelques phrases du document. D'autre part, lorsqu'on considère en particulier certaines instances comme des instances de type événement (catastrophe naturelle, ...) dont la connaissance établie peut évoluer fréquemment au cours de la période englobant la date de l'événement, les textes sur lesquels ces approches sont appliquées doivent reporter des informations nouvelles et à jour.

Dans cet article nous proposons une approche visant à identifier dans un flux de documents du Web les phrases qui reportent des informations opportunes et pertinentes sur les instances représentées d'une base de connaissances. Nous appelons ces phrases *phrases vitales*.

Détecter en temps réel les phrases vitales est une tâche complexe qui soulève plusieurs problématiques pouvant être vues comme des étapes du processus :

- comment détecter si un document reporte une information vitale sur une instance ?
- étant donné un document vital, comment extraire les phrases vitales reportant les informations vitales ?
- comment détecter si deux phrases vitales reportent la même information ?

La *première étape* est importante et nécessite la mise en place d'un processus riche que nous avons détaillé dans des travaux précédents [Abbes *et al.* (2013, 2015)]. Dans cet article, nous simplifions cette étape en nous focalisant sur des instances de type événement et en analysant le flux de documents uniquement dans les périodes de déroulement de ces événements. Par conséquent, les documents mentionnant le ou les labels associés à l'instance dans la base de connaissances ont tendance à reporter des informations vitales. La *deuxième étape* est primordiale puisqu'elle permet de choisir les phrases vitales candidates. Les travaux de l'état de l'art se fondent généralement sur la présence de mots spécifiques pour calculer un score de pertinence. Ces mots sont choisis soit manuellement, soit sélectionnés automatiquement. Dans ce travail, nous proposons d'exploiter la connaissance déjà représentée dans la base de connaissances. Nous cherchons à identifier le vocabulaire propre à chaque type d'instances. Concernant la *dernière étape*, pour détecter la nouveauté d'une phrase par rapport à une autre, nous cherchons là encore à évaluer l'apport de la connaissance déjà représentée dans la base à enrichir.

En résumé, nous souhaitons répondre à la question suivante : à quel point l'exploitation de connaissances déjà représentées peut servir à détecter les informations vitales et non redondantes relatives aux instances dans un flux de documents Web ? Nous voyons ces travaux comme une première étape qui devra ensuite être complétée par une phase d'extraction de connaissances des phrases vitales dont nous montrons le processus d'identification dans ce papier.

Cet article est organisé comme suit. Nous présentons dans la section 2 un état de l'art des travaux liés à l'identification d'informations vitales sur le Web. La section 3 présente notre approche reposant sur la prise en compte de connaissances connues sur l'instance. Dans la section 4, nous présentons et discutons l'intérêt de notre approche par rapport aux méthodes de l'état de l'art. Nous concluons et énonçons quelques perspectives en section 5.

2 État de l'art

Accélérer la mise à jour des bases de connaissances est une problématique actuelle dont le premier enjeu est d'identifier un besoin d'évolution. L'analyse de documents d'où extraire la connaissance à mettre à jour est une solution pour identifier ce besoin [Zablith *et al.* (2015)]. La phase d'identification de ces documents est souvent laissée aux concepteurs de la base dans les travaux d'ingénierie de connaissances. Cependant lorsque la base de connaissances comporte des instances largement mentionnées sur le Web, il est dommage de ne pas tirer profit de ces informations. Le module *DBpedia Live*, par exemple, vise à mettre à jour en temps quasi réel la base de connaissances lorsque les *infobox* des pages Wikipedia sont modifiées [Lehmann *et al.* (2014)]. Cependant, comme souligné dans [Frank *et al.* (2012)], un certain temps de latence est constaté pour la mise à jour alors que l'information est publiée en temps réel sur le Web.

Pour faire face à ce problème et aider à l'identification d'informations vitales lorsqu'un document de référence régulièrement mis à jour n'est pas disponible, nous nous sommes tournés vers les travaux de recherche d'information qui s'intéressent à ces aspects. La campagne

d'évaluation TREC a notamment lancé la tâche *Knowledge Base Acceleration (KBA)* [Frank *et al.* (2013, 2012)]. Plusieurs méthodes ont été proposées [Wang *et al.* (2013)] [Abbes *et al.* (2013)] [Abbes *et al.* (2015)] afin d'identifier, en temps réel, les documents vitaux reportant des informations nouvelles sur des instances de type personnes, organisations et établissements. Cependant, malgré leur utilité, ces méthodes renvoient des documents entiers ce qui oblige les éditeurs de la base de connaissances ou les outils d'extraction à parcourir tous leurs contenus pour chercher les nouvelles informations vitales. En outre, elles ne traitent pas le problème de redondance entre les documents, c'est à dire qu'elles renvoient tous les documents apportant des informations vitales même s'ils contiennent des informations vitales redondantes.

D'autres approches se sont intéressées à identifier à partir d'un flux de documents Web, les phrases vitales relatives à des événements largement connus comme les catastrophes naturelles [Aslam *et al.* (2013)] [McCreadie *et al.* (2014)]. Liu *et al.* (2013) s'appuient sur des données d'apprentissage pour apprendre les mots importants permettant d'identifier les phrases vitales. Xu *et al.* (2013) utilisent un classifieur afin de détecter les phrases contenant de nouvelles informations. Zhang *et al.* (2013) sélectionnent les phrases contenant les mots les plus représentatifs selon leur fréquence d'occurrences. Dans ce travail, nous nous intéressons aussi aux instances d'événements. Nous détectons les phrases vitales d'un nouvel événement émergent en exploitant des mots importants récupérés à partir des connaissances déjà représentées dans la base.

3 Détection en temps réel des informations vitales relatives à une instance

Notre approche a pour but de détecter en temps réel les phrases reportant de nouvelles informations vitales (pertinentes et opportunes) relatives à une instance donnée d'une base de connaissances à partir d'un flux de documents issus du Web. Ces phrases vitales peuvent servir à mettre à jour la connaissance sur cette instance. Par conséquent, elles doivent être pertinentes (concerner l'instance), exhaustives (couvrir les différentes informations publiées sur l'instance), non redondantes (reportées une seule fois) et émises sans trop de latence.

Formellement, considérons un flux continu F composé de documents d ayant chacun une date de publication $t(d)$ et une séquence de phrases s_j tels que $0 \leq j < l(d)$ où $l(d)$ désigne la longueur du document d en nombre de phrases. Soient $h_0, h_1, \ldots, h_n$ des instants séparés par un intervalle de temps constant (par exemple une heure). Nous désignons par F_{h_i} l'ensemble de documents du flux tel que $\forall d \in F_{h_i}, h_{i-1} \leq t(d) < h_i$.

L'algorithme 1 décrit le fonctionnement général de notre approche de détection des phrases vitales relatives à une instance donnée I. A chaque instant h_i, nous distinguons 3 étapes principales que nous détaillons dans les sous-sections suivantes :

1. sélection des documents vitaux D_{h_i} par rapport à I en utilisant comme requête le ou les labels associés à l'instance dans la base de connaissances,
2. sélection des phrases vitales candidates (contenant une information vitale),
3. vérification de la nouveauté des phrases candidates par rapport aux phrases déjà sélectionnées ($\in V(I)$).

Algorithm 1 Détection des phrases vitales relatives à une instance

ENTRÉES: F : Flux de documents
ENTRÉES: I : Instance à mettre à jour, ayant une étiquette *I.label*
ENTRÉES: h_0 (h_n) : Début (Fin) de la période d'analyse du flux
SORTIE: $V(I) \leftarrow \{\}$: Historique des phrases vitales relatives à I

1: **pour chaque** $i \in [1, n]$ **faire**
2: $D_{h_i} \leftarrow sélection_des_documents(F_{h_i}, I.label)$
3: **pour chaque** $d \in D_{h_i}$ **faire**
4: **pour chaque** $s_j \in d$ **faire**
5: **si** $est_vitale(s_j, I)$ ***ET*** $est_nouvelle(s_j, V(I))$ **alors**
6: $enrichir(V(I), s_j)$
7: **fin si**
8: **fin pour**
9: **fin pour**
10: **fin pour**

3.1 Sélection des documents vitaux

Nous n'analysons que la période "chaude" durant laquelle les informations vitales sur une instance donnée sont publiées dans le flux de documents du Web. Dans ce travail, nous supposons que cette période est connue. A chaque instant h_i, nous analysons les nouveaux documents apparus dans le flux entre h_{i-1} et h_i et nous attribuons à chacun d'eux un score de vitalité par rapport à l'instance *I* considérée. Ce score est calculé par la probabilité que le ou les termes composant le label de l'instance I soient générés par un modèle de langue probabiliste estimé à partir du document analysé [Zhai & Lafferty (2001)]. Le ***top-h*** des documents est sélectionné afin d'être analysé dans l'étape suivante.

3.2 Sélection des phrases vitales

Dans cette étape, nous analysons les phrases contenues dans les documents sélectionnés. Pour chaque phrase, nous devons décider si elle est vitale (reportant une information pertinente et opportune) par rapport à l'instance à surveiller I. Notre intuition est de considérer une phrase comme vitale si :

- elle est à proximité de l'instance I (des termes de l'étiquette de I),
- elle contient des mots "importants" relativement à l'instance I.

La proximité d'une phrase par rapport à l'instance I peut refléter sa pertinence. Une phrase mentionnant l'instance a plus de chance de parler de celle ci. Nous traduisons ainsi la proximité entre une phrase s_j et l'instance I en un score calculé selon l'équation suivante :

$$score_proximité(s_j, I) = \frac{1}{|I.label|} \sum_{t \in I.label} \sum_{dist=0}^{dmax} e^{-dist * match_s(t, s_{j+dist}, s_{j-dist})} \quad (1)$$

$I.label$ est l'étiquette décrivant l'instance I. $|I.label|$ est le nombre de mots qu'elle contient.
$match_s(t, s_x, s_y)$ est égal à 1 si t est contenu dans l'une des phrases s_x et s_y, 0 sinon.
$dmax$ est la distance maximale à considérer (calculée en nombre de phrases).

Nous considérons uniquement les phrases à proximité de l'instance I en favorisant celles qui sont proches de l'ensemble des termes composant l'étiquette de l'instance I, c.à.d, ayant un *score proximité* supérieur à un seuil τ_p (la valeur de τ_p peut être déterminée expérimentalement).

En plus de la proximité, nous supposons que pour une instance I, il existe un ensemble de mots "importants" qui peuvent refléter la vitalité d'une phrase. Nous appelons ces mots des *mots déclencheurs*. Nous posons l'hypothèse selon laquelle les instances de même type (représenté dans la base de connaissances) partagent les mêmes mots déclencheurs. Afin d'identifier ces mots déclencheurs, nous proposons d'exploiter toutes les annotations (description en langage naturel) qui ont pu être renseignées sur des instances du type considérées. Nous considérons comme étant une annotation le texte associé à une instance par les propriétés d'annotation de OWL, ou les propriétés du Dublin Core, ou encore le résumé associé dans DBpedia par la propriété dbpedia-owl :abstract. Par exemple, les mots tels que *effets*, *force*, *tempête*, *blessés*, *dommages* pourront être très utiles pour décrire les instances de type *ouragan* comme ils sont présents dans les annotations associées aux instances *ouragan Sandy* et *ouragan Isaac*.

Formellement, soient $X(I) = \{A(I_1), A(I_2), \ldots, A(I_m)\}$ l'ensemble des m extraits des valeurs des annotations associées aux instances de même type que I. Nous pondérons les mots t par l'équation suivante :

$$\omega(t) = \frac{\sum_{i=1}^{m} TF(t, A(I_i))}{IIF(t)} \tag{2}$$

$TF(t, A(I_i))$ est le nombre d'occurrences du terme t dans l'annotation $A(I_i)$

$IIF(t) = log(\frac{m+1}{IF(t)})$ est un facteur utilisé pour donner la priorité aux termes se trouvant dans la plupart des annotations des instances de même type que l'instance I

$IF(t)$ est le nombre d'instances du type dont l'annotation contient le terme t

Les ***top-k*** premiers mots seront considérés comme des mots déclencheurs pour l'instance I. Pour qu'une phrase soit considérée comme une phrase vitale candidate, il faut :

- que le score de proximité de la phrase soit $> \tau_p$,
- qu'elle contienne un mot déclencheur.

3.3 Détection de la nouveauté

Les phrases sélectionnées à l'étape précédente pourraient contenir des informations vitales redondantes déjà émises. Afin d'éliminer la redondance, nous comparons chaque phrase vitale candidate à toutes les phrases vitales déjà ajoutées à l'ensemble incrémental $V(I)$. Détecter la nouveauté n'est pas une tâche facile. Comme le montre le tableau 1, les deux phrases s1 et s2 contiennent un grand nombre de chaînes de caractères en commun, mais reportent deux informations différentes. Inversement, les phrases s2 et s3 sont divergentes textuellement mais portent la même information.

Dans notre approche, nous considérons qu'une phrase vitale candidate s_j est nouvelle par rapport aux phrases déjà émises $V(I)$ si son texte est divergent et/ou présente une instance liée nouvelle (non détectée dans les phrases précédentes $V(I)$). Formellement, s_j est nouvelle si elle respecte la fonction de nouveauté suivante :

n°	Date	Texte
s1	26 Oct. 2012 - 07 :27	Hurricane Sandy leaves 21 people dead in Caribbean
s2	26 Oct. 2012 - 20 :50	Hurricane Sandy leaves 41 people dead in Caribbean
s3	30 Oct. 2012 - 06 :17	Hurricane Sandy is continuing to head north from the Caribbean where it has killed a total of 41 people in the Caribbean

TABLE 1 – Exemple de phrases vitales

$$est_nouvelle(s_j, V(I)) = \textit{texte_divergent}(s_j, V(I)) \circ \textit{instance_liée_nouvelle}(s_j, V(I)) \quad (3)$$

$$texte_divergent(s_j, V(I)) = \begin{cases} faux & si\ \exists s_k \in V(I),\ cos(s_j, s_k) > \tau_n(V(I)) \\ vrai & sinon \end{cases} \quad (4)$$

$$instance_liée_nouvelle(s_j, V(I)) = \begin{cases} vrai & si\ \exists x \in \textit{IL}(s_j, I),\ \forall s_k \in V(I)\ x \notin \textit{IL}(s_k, I) \\ faux & sinon \end{cases} \quad (5)$$

$IL(s_i, I)$ est l'ensemble des instances liées reconnues dans la phrase s_i. Dans notre méthode, nous proposons de prendre en compte les propriétés définies dans l'ontologie pour le concept type de l'instance I que nous considérons. Nous cherchons à identifier dans la phrase, des instances ou des valeurs potentiellement liées sémantiquement car leur type correspond au domaine ou co-domaine des propriétés définies dans l'ontologie pour le concept.

$\tau_n(V)$ est un seuil de nouveauté textuelle. Au fur et à mesure que l'ensemble de phrases vitales $V(I)$ s'enrichit, le risque de redondance augmente, d'où l'idée de faire décroître le seuil τ_n selon une fonction gaussienne :

$$\tau_n(V(I)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{|V(I)|^2}{\delta^2}} \quad (6)$$

Les paramètre σ a un impact sur la tolérance de la similarité, et le paramètre δ contrôle le taux de décroissance du seuil. $|V(I)|$ est le nombre de phrases de l'ensemble $V(I)$.

Le symbole $\circ$ de l'équation 3 peut être un opérateur ***ET*** pour rendre le système orienté Précision en limitant la redondance (dans ce cas, la phrase s3, ne présentant aucune instance liée nouvelle par rapport aux phrases s1 et s2, sera considérée comme redondante malgré que le fait qu son texte diverge), ou bien un opérateur ***OU*** pour privilégier le Rappel (dans ce cas, la phrase s2, malgré le fait qu'elle diverge peu par rapport à s1, elle sera considérée comme nouvelle car elle présente une nouvelle valeur).

4 Expérimentations

Trouver un cadre expérimental pour évaluer notre approche n'est pas trivial car il n'existe pas à notre connaissance de jeux de données constitués d'une base de connaissances dont plusieurs versions sont disponibles ainsi que les corpus de documents desquels la connaissance a été extraite pour mettre en place les différentes versions. Bien qu'étant une tâche de Recherche d'Information, la tâche *Temporal Summarization* (TS) de la campagne d'évaluation TREC est à notre sens la plus adaptée.

ID	Ressource dans dbpedia	Requête (label)	Type	Début	Fin
1	2012_Buenos_Aires_rail_disaster	buenos aires train crash	accident	2012-02-22-12	2012-03-03-11
9	2012_Guatemala_earthquake	guatemala earthquake	earthquake	2012-11-07-16	2012-11-17-16
19	2012_Romanian_protests	romanian protests	protest	2012-01-12-00	2012-01-26-00

TABLE 2 – Exemples d'instances proposées dans la tâche TS en 2013 et 2014

4.1 Cadre expérimental

Le but de cette tâche est de concevoir des systèmes capables de surveiller les événements en détectant à la volée toutes les nouvelles informations publiées dans un flux qui comporte 500 millions de documents en anglais, issus de différentes sources du Web (Presse, Blog, etc.) et associés à des dates de publications (*timestamp*) allant du mois d'octobre 2011 au mois d'avril 2013. Les systèmes doivent extraire les phrases contenant des informations vitales tout en évitant la redondance.

4.1.1 Instances

Les topics à considérer dans le cadre de cette tâche correspondent à des événements d'actualité tels que des manifestations, des accidents ou des catastrophes naturelles. Le tableau 2 illustre quelques exemples parmi les 24 événements [1] proposés par les organisateurs de la tâche en 2013 et 2014. Les colonnes *Début* et *Fin* définissent la période à surveiller pour chaque événement. En analysant ces topics, nous avons remarqué qu'ils correspondaient à des instances de DBpedia pour lesquelles un label est défini. Nous considérons que les instances pour lesquelles nous souhaitons identifier des phrases vitales dans notre approche peuvent être apparentées aux topics. Le label défini dans DBpedia est utilisé pour constituer la requête dans notre approche. Il est présenté dans la troisième colonne du tableau.

4.1.2 Informations vitales à retrouver et jugements de pertinence

La figure 1 illustre le nombre d'informations vitales à retrouver pour les instances proposées. Ces informations sont extraites à partir des différentes mises à jour des pages Wikipedia de ces événements. Ces informations ayant été rajoutées manuellement au cours d'une mise à jour de la page de Wikipedia, nous considérons qu'elles auraient également mené à une mise à jour manuelle de la connaissance représentée dans DBpedia sur l'instance. Cette référence nous paraît donc pertinente pour évaluer notre approche.

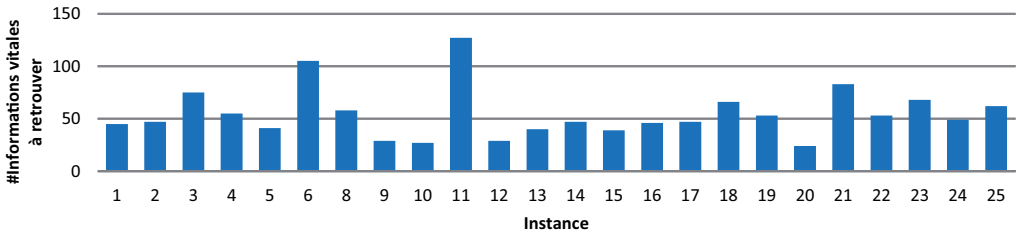

FIGURE 1 – Nombre d'informations vitales à retrouver pour les instances proposées dans la tâche TS en 2013 et 2014

1. Les *topics* sont disponibles dans ce lien *www.trec-ts.org/documents*

Une phrase est jugée pertinente si elle peut être associée à au moins une information vitale. Cette association est faite par les juges de la tâche. Dans l'exemple de l'instance *2012_Buenos_Aires_rail_disaster*, la phrase *49 dead, over 500 wounded in Buenos Aires !*, émise le *23-02-2012 ; 03 :21*, est associée à trois informations vitales : *"train accident in Buenos Aires, Argentina"*, *"550 injured"* et *"49 confirmed deaths"*.

4.1.3 Mesures d'évaluation

Pour analyser les résultats, nous utilisons les mesures suivantes classiques de Rappel et Précision, ainsi que :

$$Précision_N = \frac{Nombre\ d'informations\ vitales\ détectées}{Nombre\ total\ de\ phrases\ émises} \tag{7}$$

$$H = 2 * \frac{Précision * Rappel}{Précision + Rappel} \tag{8}$$

$$H_N = 2 * \frac{Précision_N * Rappel}{Précision_N + Rappel} \tag{9}$$

Nous utilisons le rappel, la precision et (8) pour mesurer la capacité d'un système à renvoyer les phrases vitales, sans pénaliser la redondance. Pour considérer la nouveauté (pénaliser la redondance), nous utilisons le rappel, (7) et (9).

4.2 Configuration de notre système

Pour l'étape de ***sélection des phrases vitales candidates***, nous appliquons la méthode expliquée dans la section 3.2 qui repose sur la détection du ***topK*** des mots déclencheurs relatif à l'instance traitée (Eq. 2). Nous désignons cette stratégie par ***Gen-Auto***.
Nous évaluons aussi deux autres stratégies :

- La stratégie ***Gen-Man*** : sélection manuelle de 15 mots génériques qui peuvent caractériser tous ou la plupart des événements évalués. Parmi ces mots-clés, nous listons : *died, dead, death, kill, injuries, damage, victims, survivor* etc.
- La stratégie ***QW*** : considération des termes du label de l'instance etudiée comme mots déclencheurs.

Pour la ***détection de la nouveauté***, nous évaluons les méthodes suivantes :

- ***Texte*** : utilisant uniquement la nouveauté textuelle (Eq. 4)
- ***NER*** : utilisant uniquement la reconnaissance d'instances liées (Eq. 5). Les instances considérées dans la tâche sont du type Event. Puisque les propriétés [2] définies pour ce concept dans l'ontologie DBpedia sont très discutables, nous avons choisi d'exploiter la définition de l'ontologie Event [3] en recherchant dans les phrases des instances de lieux, de personnes et des valeurs numériques [4].
- ***NER*Texte*** : utilisant la fonction de nouveauté combinée avec un opérateur ET (Eq. 3)
- ***NER+Texte*** : utilisant la fonction de nouveauté combinée avec un opérateur OU (Eq. 3)
- ***Sans*** : Sans appliquer la nouveauté (émettre toutes les phrases sélectionnées à l'étape 2)

2. http ://mappings.dbpedia.org/server/ontology/classes/Event
3. http ://motools.sourceforge.net/event/event.html
4. nous utilisons l'outil réalisé par le groupe *NER Stanford* (http ://nlp.stanford.edu/ner/)

Paramètres de notre système

Nous avons appliqué la validation croisée afin de fixer les paramètres de notre système, en faisant varier le nombre de documents sélectionnés par heure entre 1 et 20 avec un pas de 1, *top-k* entre 4 et 40 avec un pas de 2, τ_p entre 0.4 et 1 avec un pas de 0.1, δ entre 10 et 300 avec un pas de 10, σ entre 0.5 et 1 avec un pas de 0.1. Les valeurs optimales obtenues sont : *top-h=10* et *top-k=15*, $\tau_p = 0.8$, $\delta = 200$ et $\sigma = 0.5$.

4.3 Analyses des résultats

A l'issue de la première étape, notre système renvoie $20\,800$ documents pour les 24 instances (soit 866 documents par instance) permettant d'atteindre un rappel moyen de 0.65.

4.3.1 Stratégies de sélection des mots déclencheurs

La figure 2 compare les différentes stratégies de sélection des mots déclencheurs pour la détection des phrases vitales sans tenir compte de la redondance (***sans***). Considérer comme mots déclencheurs uniquement les termes du label associé à l'instance (***QW***) permet de capturer environ **63%** (0.407/0.650) des informations vitales contenues dans les documents sélectionnés avec une précision ne dépassant pas ***0.161***. La condition de proximité (Eq. 1) avec un seuil $\tau_p = 0.8$ semble être stricte car elle exige la présence de la plupart des termes du label dans les phrases vitales ce qui peut expliquer la perte de 37% d'informations vitales. L'utilisation des mots-clés **Gen-Auto** revient à vérifier la présence simultanée des termes du label et d'un mot générique. Comme résultat, on constate une amélioration légère de la précision par rapport à QW "pratiquement" sans baisse du rappel. Cette stabilité du rappel prouve que les mots saillants récupérés automatiquement des annotations associées aux instances du même type permettent de couvrir les différents aspects de la nouvelle instance traitée. L'amélioration de la précision montre l'importance de ces mots. La sélection manuelle de mots génériques (**Gen-Man**) améliore la précision (surtout pour les instances de 2014) mais le rappel est relativement faible par rapport à la méthode automatique *Gen-Auto*.

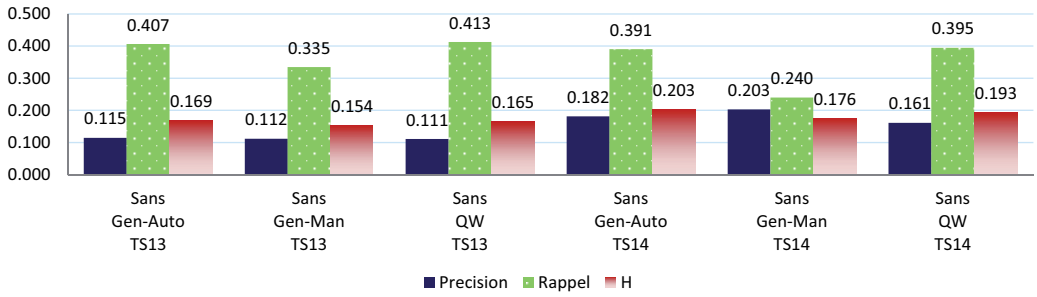

FIGURE 2 – Comparaison des différentes stratégies de sélection des mots déclencheurs (Gen_Auto, Gen_Man, QW) pour la détection des phrases vitales

4.3.2 Comparaison des différentes configurations de détection de la nouveauté

La figure 3 compare les différentes configurations de détection de la nouveauté. L'application du module de nouveauté permet d'améliorer la $precision_N$ en pénalisant le rappel. Combiner

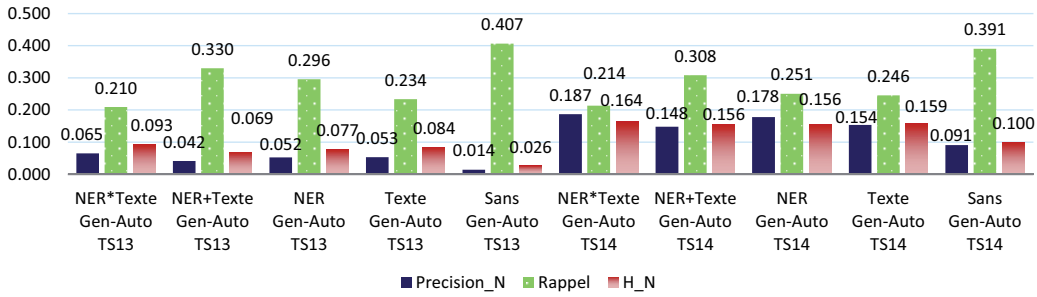

FIGURE 3 – Comparaison des différentes méthodes de détection de la nouveauté

la similarité textuelle avec la reconnaissance d'instances liées ***NER*Texte*** donne une meilleure moyenne harmonique (H_N) entre le rappel et la précision pour les instances de 2013 et 2014. Utiliser la stratégie *NER+Texte* reste utile si nous privilégions l'exhaustivité de la connaissance à extraire pour l'instance.

4.3.3 Comparaison de notre système par rapport aux systèmes participant à la tâche

Dans cette section, nous comparons notre système par rapport aux meilleurs systèmes ayant participé à la tâche en utilisant l'outil d'évaluation officiel [5] avec les jugements de pertinences officiels. Les mesures *ELG* et *LC* sont similaires aux mesures de précision et rappel respectivement mais en pénalisant la redondance et la latence lors de la détection d'informations [Aslam *et al.* (2013)]. Notre système aurait pu être classé premier (/7 participants) dans la tâche de TS 2013, et troisième (/6 participants) pour l'année 2014. Notre système apparaît donc comme efficace pour la détection de phrases vitales pour la mise à jour de bases de connaissances.

TS 2013				TS 2014			
Système	**ELG**	**LC**	**H-ts**	**Système**	**ELG**	**LC**	**H-ts**
*Gen-Auto ; Text*NER*	0.1102	0.1986	**0.1355**	cunlp	0.0631	0.322	**0.1162**
Gen-Auto ; Text+NER	0.0768	0.2619	0.1188	BJUT	0.0657	0.4088	0.1110
ICTNET	0.0794	0.3636	0.1078	*Gen-Auto ; Text*NER*	0.0881	0.1646	0.1047
PRIS	0.136	0.195	0.1029	uogTr	0.0467	0.4453	0.0986
HLTCOE	0.0522	0.2834	0.0827	*Gen-Auto ; Text+NER*	0.0712	0.2181	0.0963

TABLE 3 – Comparaison de notre système par rapport aux systèmes participant à la tâche TS 2013 et 2014. H-ts est la moyenne harmonique entre ELG et LC.

4.3.4 Rapidité de notre approche par rapport aux mises à jour de Wikipedia

La figure 4 compare la rapidité de notre système (Gen-Auto ; NER*Texte) à détecter les informations vitales pour les 24 événements par rapport aux mises à jour Wikipedia. Notre système permet de détecter 67% d'informations vitales avant que celles-ci soient mises à jour dans Wikipedia. La moitié des informations sont détectées par notre système 7 heures (au moins) avant qu'elles ne soient mises à jour dans Wikipedia. En moyenne, notre système permet de gagner 18 heures.

Dans le tableau 4, nous illustrons quelques exemples d'informations vitales détectées par notre système avant qu'elles soient ajoutées dans Wikipedia.

5. http ://www.trec-ts.org/documents

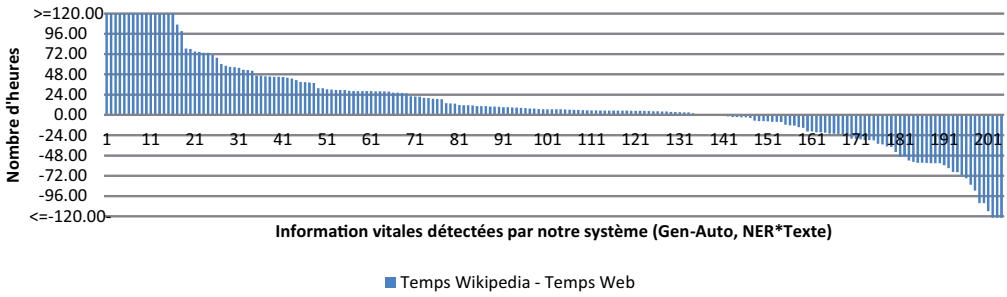

FIGURE 4 – Évaluation de la rapidité de notre système (Gen-Auto ; NER*Texte) par rapport aux mises à jours Wikipedia

Id de l'instance	*Information vitale détectée*	t_{web}	t_{wp}	t_{IB}	*Gain*
1	*550 injured*	22-02-12 16 :05	22-02-12 22 :49	22-02-12 22 :49	6.7h
1	*crashed at speed of 26 kilometers per hour*	22-02-12 22 :21	22-02-12 23 :01	-	0.67h
9	*39 casualties reported in Guatamala*	08-11-12 00 :33	08-11-12 04 :33	08-11-12 04 :33	1h
9	*48 casualties reported*	08-11-12 07 :42	08-11-12 07 :55	08-11-12 07 :55	0.22h
19	*Early modest estimates put over 5000 people in the streets of Romanian cities*	16-01-12 03 :58	18-01-12 02 :28	-	46.5h
19	*Queensland floods*	27-01-13 11 :35	24-01-13 22 :42	-	60.8h

TABLE 4 – Exemple d'informations vitales détectées par notre approche (Gen-Auto, NER*Texte). t_{web}, t_{wp}, t_{IB} représentent les temps de la disposition de l'information par notre système, dans Wikipedia et dans les infoboxes de Wikipedia respectivement.

La figure 4 et les exemples du tableau 4 montrent que les informations sont généralement publiées dans le documents Web (presse, blogs, etc.) avant qu'elles soient éditées dans les encyclopédies collaboratives comme Wikipedia. Notons que la mise à jour n'est pas forcément reportée dans les InfoBoxes principalement exploités pour enrichir DBpedia. Bien que les instances analysées représentent des événements largement connus, qui intéressent plusieurs contributeurs, nous remarquons toujours un temps de latence. Ce temps de latence devrait être plus grand pour des instances "moins connues".

5 Conclusion

Dans ce travail, nous avons proposé une méthode qui permet d'extraire les informations vitales au fur et à mesure qu'elles apparaissent dans le Web. L'importance de ce type d'approches nous paraît cruciale afin d'accélérer la mise à jour des bases de connaissances. Ces approches sont utiles non seulement pour aider à la mise à jour de documents collaboratifs décrivant des instances comme les pages wikipedia, mais aussi pour la mise à jour des bases de connaissances elles-mêmes car elles permettent d'identifier les phrases spécifiques pouvant ensuite être analysées par des extracteurs (tels que ceux décrit dans Zablith *et al.* (2015)) pour enrichir la base. L'expérimentation que nous avons menée montre que des mises à jour plus fines de DBpedia pourraient notamment être mises en oeuvre par l'identification en temps réel de phrases vitales issues du web dont l'information ne se trouve pas toujours dans l'infoBox. Nous souhaitons à très court terme poursuivre les évaluations de notre système en utilisant des outils d'extractions de connaissances sur les phrases vitales retrouvées. Nous souhaitons également trouver un autre cadre d'évaluation pour lequel des documents et une base de connaissances dans laquelle des connaissances plus formellement représentées sont disponibles.

Références

ABBES R., PINEL-SAUVAGNAT K., HERNANDEZ N. & BOUGHANEM M. (2013). IRIT at trec knowledge base acceleration 2013 : Cumulative citation recommendation task. In *Proceedings of the Twenty-Second Text REtrieval Conference*, Gaithersburgh, MD, USA.

ABBES R., PINEL-SAUVAGNAT K., HERNANDEZ N. & BOUGHANEM M. (2015). Leveraging temporal expressions to filter vital documents related to an entity. In *ACM Symposium on Applied Computing (SAC) (to appear)*.

ASLAM J., DIAZ F., EKSTRAND-ABUEG M., PAVLU V. & SAKAI T. (2013). Trec 2013 temporal summarization. In *Proceedings of the Twenty-Second Text REtrieval Conference*, Gaithersburgh, USA.

AUGENSTEIN I., PADÓ S. & RUDOLPH S. (2012). Lodifier : Generating linked data from unstructured text. In *Proceedings of the 9th International Conference on The Semantic Web : Research and Applications*, ESWC'12, p. 210–224, Berlin, Heidelberg.

EXNER P. & NUGUES P. (2012). Entity extraction : From unstructured text to dbpedia rdf triples. WoLE'12.

FRANK J. R., BAUER S. J., KLEIMAN-WEINER M., ROBERTS D. A., TRIPURANENI N., ZHANG C. & RE C. (2013). Evaluating stream filtering for entity profile updates for trec 2013. In *Proceedings of the Twenty-Second Text REtrieval Conference*, Gaithersburgh, MD, USA.

FRANK J. R., KLEIMAN-WEINER M., ROBERTS D. A., NIU F., ZHANG C., RE C. & SOBOROFF I. (2012). Building an Entity-Centric stream filtering test collection for TREC 2012. In *Proceedings of the Text REtrieval Conference (TREC)*.

LEHMANN J., ISELE R., JAKOB M., JENTZSCH A., KONTOKOSTAS D., MENDES P., HELLMANN S., MORSEY M., VAN KLEEF P., AUER S. & BIZER C. (2014). DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*.

LIU Q., LIU Y., WU D. & XUEQI C. (2013). Ictnet at temporal summarization track trec 2013. In *Proceedings of the Twenty-Second Text REtrieval Conference*, Gaithersburgh, MD, USA.

MCCREADIE R., MACDONALD C. & OUNIS I. (2014). Incremental update summarization : Adaptive sentence selection based on prevalence and novelty. In *Proceedings of the 23rd International Conference on Conference on Information and Knowledge Management*, p. 301–310, New York, USA.

PETASIS G., KARKALETSIS V., PALIOURAS G., KRITHARA A. & ZAVITSANOS E. (2011). Ontology population and enrichment : State of the art. In *Knowledge-driven multimedia information extraction and ontology evolution*, p. 134–166 : Springer-Verlag.

WANG J., SONG D., LIN C.-Y. & LIAO L. (2013). BIT and MSRA at trec kba ccr track 2013. In *Notebook of the TExt Retrieval Conference 2013 (TREC 2013)*, Gaithersburgh, MD, USA.

XU T., MCNAMEE P. & W.QARD D. (2013). HLTCOE at TREC 2013 : Temporal summarization. In *Proceedings of the Text REtrieval Conference*, Gaithersburgh.

ZABLITH F., ANTONIOU G., D'AQUIN M., FLOURIS G., KONDYLAKIS H., MOTTA E., PLEXOUSAKIS D. & SABOU M. (2015). Ontology evolution : a process-centric survey. *The Knowledge Engineering Review*, **30**(01), 45–75.

ZHAI C. & LAFFERTY J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, p. 334–342, New York, NY, USA.

ZHANG C., XU W., MENG F., LI H., WU T. & XU L. (2013). The information extracion systems of pris at temporal summarization track. In *Proceedings of the Text REtrieval Conference*, Gaithersburgh.