Évolution des profils d'entités à l'aide d'un modèle de langue sensible au temps

Vincent Bouvier[*,**] — **Patrice Bellot**[**]

[*] *Kware, 565 Rue Marcelin Berthelot, Aix-en-Provence, France*
[**] *Aix-Marseille Université, CNRS, LSIS UMR 7296, Marseille, France*

RÉSUMÉ. *Retrouver des informations importantes en temps sur une entité nommée particulière est un réel challenge. En effet, cela implique d'être capable de détecter l'entité dans les documents, mais en plus d'être capable de qualifier d'importante, au regard de l'entité, l'information véhiculée par le document. Dans cet article, nous formalisons un modèle de langue sensible au temps, et nous l'utilisons dans les profils d'entités. Nous mettons en place un ensemble de méta critères qui utilisent pleinement l'amélioration du profil d'entité. L'utilisation de méta critères nous assure d'avoir un système non dépendant de l'entité et donc place notre approche dans le domaine du semi-supervisé. Nous évaluons notre approche sur les données de la campagne d'évaluation TREC sur la tâche KBA 2013 et nous obtenons de bonnes performances et des conclusions intéressantes.*

ABSTRACT. *Finding important information in real time on a particular named entity is a real challenge. It requires to be able to detect the entity within the document and to be able to assess how important the information is regarding the entity. In this article, we formalize a new time-aware langue model that we use as part of entity profiles. We design meta criteria to fully use this new profile design. Using meta criteria ensure to have a entity independent system and make our approach semi supervised. We evaluate our approach on the data from the TREC on the KBA 2013 track and we obtain satisfying results and interesting conclusions.*

MOTS-CLÉS : *entités nommées, modèle de langue sensible au temps, filtrage de documents, classification, méta critères.*

KEYWORDS: *named entities, time-aware language model, document filtering, classification, meta criteria.*

1. Introduction

La plupart du temps, les systèmes de recherche d'information sont utilisés à l'aide d'un ensemble de mots clés (requête) qui définissent un besoin. Le système de recherche d'information a pour but de retrouver et d'ordonner les documents du plus pertinent au moins pertinents. Cependant, il n'est pas toujours évident d'exprimer sous forme de requête un besoin. Par exemple, pour la recherche d'entités nommées (personnes, organisations, entreprises...), la tâche est plus complexe. La description d'une entité peut se faire par un nom (une forme de surface), mais ce n'est pas suffisant. Certaines entités partagent le même nom. Il faut alors être capable de différencier les entités homonymiques. Par ailleurs, certaines entités peuvent évoluer au cours du temps et par voie de conséquence le champ lexical autour de l'entité peut également changer. La problématique de suivi d'entité nommée s'attache à filtrer, sur un flux de documents continu, les documents dits vitaux pour une entité. Un document vital est un document qui contient, soit une information nouvelle, soit une information importante concernant dans l'entité. La détection de la vitalité d'un document implique le fait de suivre l'évolution d'une entité en s'assurant que les informations sur l'entité ne dérivent pas du sujet principal.

La recherche d'information centrée sur les entités nommées trouve son utilité dans les domaines d'applications de la veille informationnelle ou encore dans le domaine du marketing. (Frank *et al.*, 2012) montrent également qu'il est difficile de maintenir à jour des bases de connaissance (tel que Wikipédia) du fait de leurs grandes envergures. Ils montrent que pour certaines pages centrées sur des entités peu populaires, le temps delta médian entre le moment où une nouvelle apparaît sur internet et le moment où la nouvelle fait l'objet d'une mise à jour de l'article est de 365 jours. Dans ces domaines, la réactivité, la détection d'évènements, de nouveauté sont des atouts essentiels. Les systèmes de filtrage d'information sont alors plus appropriés pour ce type de tâche. Comme le définissent (Belkin et Croft, 1992), un système de filtrage d'information doit sélectionner un sous ensemble de données issues d'un flux de données continu. Dans le cas de recherche de documents sur le web, le flux est composé de documents textuels structurés au format standard HTML. Chacun des documents doit alors être analysé pour savoir s'il contient des informations d'intérêt concernant l'entité recherchée. Cela implique d'avoir *a priori* plusieurs éléments pour faire cette analyse : 1. des connaissances concernant l'entité recherchée ; 2. des indicateurs qui permettent la détection de nouveauté, d'importance du document.

Dans cet article, nous proposons une approche semi-supervisée qui ambitionne de filtrer les documents qui contiennent une information importante à propos d'entités définies. Nous utilisons une approche statistique qui permet d'évaluer des documents concernant des entités pour lesquels nous n'avons aucune information annotée ce qui rend notre méthode indépendante des entités. Nous testons notre approche sur les données issues de la campagne d'évaluation TREC pour la tâche Knowledge Base Accelaration (KBA) de 2013 et nous présentons des résultats aussi bons voir meilleurs que ceux présentés pour la conférence.

2. Travaux connexes

Le challenge que nous proposons de relever a également été porté depuis 2012 par la tâche Knowledge Base Acceleration (Frank *et al.*, 2012) issue de la campagne d'évaluation TREC (Text REtrieval Conference). Cette tâche se concentre sur le filtrage de documents vitaux centrés sur les entités nommées. Un document est dit vital pour une entité lorsque celui-ci contient une information qui affecte particulièrement l'entité. Cette tâche trouve de nombreuses applications de la veille informationnelle à la mise à jour automatique de bases de connaissance comme Wikipedia. L'étude de (Frank *et al.*, 2012) montre qu'il peut exister un temps de latence considérable de 365 jours entre le moment où une information est publiée sur internet et le moment où cette information est ajoutée dans l'article Wikipedia concerné.

Il est possible de décrire une entité en utilisant un ensemble de mots clés qui souvent font référence au moyen de mentionner l'entité dans un document. Cependant, une entité peut être citée de différentes manières (par exemple : Elvis, The King). Nous appelons les différentes manières de mentionner une entité *Variantes de nom*. De plus, deux entités peuvent partager un même nom (Boris Berezovsky, l'homme d'affaires et le pianiste). L'utilisation de simples mots clés peut alors être problématique. (Navigli, 2009) montrent dans leur étude que l'utilisation du contexte peut permettre de désambiguïser les mots. Nous pensons qu'il en est de même pour les entités. Si nous reprenons l'exemple de Boris Berezovsky, le vocabulaire utilisé dans le document faisant référence à l'homme d'affaires sera *a priori* très différent de celui faisant référence au pianiste.

(Sehgal et Srinivasan, 2007) assimilent un profil d'entité comme étant un modèle de langue. Ils construisent un modèle de langue basé sur les top-n documents obtenus après avoir soumis une requête au moteur de recherche Google. La requête était composée du nom de l'entité et de variantes de nom trouvés manuellement. Ils ont évalué leur approche en comparant le modèle de langue obtenu avec la page Wikipedia de l'entité en utilisant une mesure de similarité. Les résultats montrent une forte similarité entre les deux modèles de langue. Pour les entités pour lesquels aucune page Wikipedia n'est disponible, cette méthode peut être relativement efficace. Pour les autres entités, la page Wikipedia peut suffire à constituer le modèle de langue. En complément, (Cucerzan, 2007) proposent une méthode qui, à partir d'une page Wikipedia centrée sur une entité, extrait d'autres informations à l'aide d'heuristiques et de parcours du graphe de connaissance. Les auteurs ont remarqué que très souvent les mots en gras situés dans le premier paragraphe de l'article, correspondent à une variante de nom. En explorant le graphe de connaissance, il est possible de connaître toutes les pages ayant un lien qui pointe vers la page de l'entité et vice et versa. Enfin, lorsqu'un lien est détecté depuis une page vers la page de l'entité, la légende du lien peut être utilisée comme variante de nom.

(Efron, 2014) proposent une méthode pour mettre à jour le modèle de langue associé à une entité. Cependant, la méthode proposée provoque une baisse de performance du système. En effet, mettre à jour un modèle de langue peut s'avérer

risqué puisque le modèle peut éventuellement dériver du sujet principal. Notre approche propose également d'utiliser les documents pour la mise à jour du profil. Afin d'éviter toute dérive, nous avons séparé le modèle de langue qui décrit l'entité de celui qui reflète ce qui se passe pour l'entité. Le premier modèle de langue est le modèle de référence, quant au second, c'est le modèle *time-aware* qui prend en considération à la fois les nouveaux documents, mais également la date de publication de ces derniers pour estimer son modèle.

(Liu et Fang, 2013) définissent un profil d'entité comme un ensemble de liens entre entités. Ils définissent une fonction de score et testent de manière empirique différents paramètres afin d'obtenir le meilleur score sur la campagne KBA en 2012. Cette méthode ne permet cependant pas de considérer d'autres facteurs tels que la nouveauté, ou encore l'impact de l'information sur le Web. (Kjersten et McNamee, 2012) utilisent un classifieur de type Machine à Vecteur de Support (SVM) et obtiennent le deuxième meilleur score de 2012. Leur approche est complètement supervisée (puisqu'ils construisent un modèle par entité) et donc nécessite de nouvelles informations d'entraînement pour chaque nouvelle entité. Cela représente une très grande contrainte. Nous mettons en place une approche dite semi-supervisée qui utilise une approche statistique sur les données d'entraînement disponibles pour déduire des règles qui seront ensuite utilisées pour toutes les entités. Notre système est donc indépendant des entités sur lesquelles il s'entraîne. Durant KBA 2013, le système présenté par (Bellogín et Gebremeskel, 2014) utilise deux classifieurs en cascade, comme introduit par(Bonnefoy *et al.*, 2013a), avec un ensemble de méta critères inspirés des études de (Balog *et al.*, 2013 ; Bonnefoy *et al.*, 2013b).

Dans les études précédentes, les méta critères utilisés étaient essentiellement focalisés sur la désambiguïsation de l'entité dans un document. Cependant, ces méta critères ne prennent pas en compte les changements que peut subir une entité. Pour caractériser la nouveauté, il est possible d'utiliser des mesures de similarité/divergence telles que la similarité cosine ou la divergence de Jensen-Shannon (Endres et Schindelin, 2003). Ces mesures permettent d'estimer une distance entre deux vecteurs. Il est possible par exemple d'utiliser ces mesures entre un modèle de langue qui représente les connaissances *a priori* connues sur l'entité, et un nouveau document. (Karkali *et al.*, 2014) ont testé différentes approches pour mesurer la nouveauté sur des données du monde réel. Ils montrent que le score de nouveauté donnée par une comparaison entre un document de référence et le résumé du document jugé offre des performances supérieures à la plupart des approches classiques en plus d'être rapide. Le score de nouveauté est calculé à l'aide d'une version lissée du $tf \cdot idf$ avec une notion de temporalité. Cette approche est très intéressante, mais elle requiert cependant une structure adaptée qui permet de garder une trace non seulement des documents, mais également de la date associée aux documents.

Nous proposons dans cet article une représentation améliorée du profil d'entité qui permet de construire un modèle de langue sensible au temps sans altérer complètement l'entité afin d'éviter une dérive du profil. Nous formalisons complètement ce modèle de langue temporel (*Time-Aware Language Model* TALM). (Li et Croft, 2003) a déjà

introduit une notion de modèle de langue sensible au temps. Ce dernier utilise le temps pour donner plus d'importance à un document publié récemment. Notre approche ambitionne d'utiliser le temps comme une manière d'oublier les informations au fur et à mesure que le temps passe pour se consacrer uniquement sur ce qui se passe à l'instant t pour une entité. Nous introduisons également de nouveaux méta critères qui se basent sur ce nouveau profil. Ils ont pour but d'aider à désambiguïser une entité dans un document, découvrir l'impact du document sur le Web et mesurer le degré de nouveauté apportée pour l'entité recherchée dans le document.

3. Utilisation de deux modèles de langue dans un profil d'entité

Nous avons appris des travaux de (Efron, 2014) que la mise à jour de profil d'entité est délicate. En effet, si les documents proviennent d'un système d'évaluation automatique il est possible que le profil dérive. Puisque nous utilisons ce genre de système, nous proposons alors d'avoir deux modèles de langue dans un profil d'entité. Chacun de ces deux modèles sert des buts différents :

- Le modèle de langue de Référence (*Reference Language Model, RLM*) estime des probabilités sur les mots en se basant sur un ensemble de documents qui regroupent des informations fondamentales (biographiques) sur l'entité. La constitution d'un tel ensemble de documents peut être problématique puisqu'elle doit être fiable (annotée manuellement). Il est tout à fait possible d'utiliser par exemple l'article Wikipédia dédié à l'entité lorsque ce dernier existe. Par ailleurs, il n'est pas nécessaire d'avoir un grand nombre de documents. C'est donc une contrainte qui est très minime.

- Le modèle de langue sensible au temps (*Time-Aware Language Model, TALM*) permet d'estimer des probabilités sur les mots en se basant deux aspects : 1. un ensemble de documents trouvés de manière automatique 2. la date de publication de chacun des documents. Le but de ce modèle est de capturer en temps réel les informations d'actualités pour une entité.

3.1. *Le modèle de langue de Référence (RLM)*

Le RLM est associé à un ensemble de documents D. Soit $tf(w, d)$ la fonction qui donne le nombre d'occurrences d'un mot w dans le document d. La fonction $tf(w, D)$ donne le nombre d'occurrences d'un mot w dans l'ensemble des documents D :

$$tf(w, D) \quad = \quad \sum_{d \in D} tf(w, d) \qquad [1]$$

Soit $|d|$ le nombre total d'occurrences de mots dans le document d. La probabilité d'apparition d'un mot w dans un ensemble de documents D est définie par :

$$p(w|D) \quad = \quad \frac{\sum_{d \in D} tf(w,d)}{\sum_{d \in D} |d|} \qquad [2]$$

3.2. *Le modèle de langue sensible au temps (TALM)*

Pour capturer les informations concernant une entité sans risquer d'altérer le modèle de langue de référence nous utilisons un deuxième modèle de langue sensible au temps (TALM). Afin d'éviter toute dérive thématique qui éloignerait trop le TALM de l'entité cible, la probabilité d'apparition d'un mot dépend de deux facteurs : 1. la fréquence d'apparition du mot jusqu'a l'instant présent ; 2. la pondération donnée par une fonction de déclin Δ. Pour chaque instant $\{t-k, ..., t-2, t-1, t\}$, la fréquence d'apparition est pondérée à l'aide d'une fonction de déclin qui représente l'oubli d'un mot. Plus la date d'apparition du mot est éloignée de l'instant t, plus le facteur de déclin est proche de 0 (et donc la probabilité d'apparition du mot est proche de 0). La fonction de déclin dépend de deux paramètres. Le premier paramètre ρ permet de régler les bornes début et fin du déclin de manière symétrique. Le paramètre temporel λ définit la durée totale du déclin (combien de temps s'écoule entre le moment où $x = 0$ et $x = 1$), il s'exprime dans la même unité que les instants t. À des fins d'optimisation mémoire, nous arrondissons les valeurs de Δ au-delà des bornes $]0; 1[$.

Considérons deux instants t_{e1} et t_{e2} avec $t_{e1} \geq t_{e2}$, nous définissons les fonctions δ et Δ (équation 3 et 4) qui permettent l'oubli progressif des mots (la baisse de la probabilité qui leur est associée dans le modèle de langue dynamique TALM) tel que :

$$\begin{array}{rccl} \delta := & \mathbb{R}^2 & \rightarrow & \mathbb{R} \\ & (t_{e1}, t_{e2}) & \mapsto & \frac{1}{\lambda} \cdot (t_{e1} - t_{e2}) \end{array} \quad [3]$$

$$\Delta(t_{e1}, t_{e2}) = \begin{cases} 1, & \text{si } \delta(t_{e1}, t_{e2}) = 0 \\ 0, & \text{si } \delta(t_{e1}, t_{e2}) \geq 1 \\ 1/(1 + e^{\rho(\delta(t_{e1}, t_{e2}) - 0.5)}) & \text{sinon} \end{cases} \quad [4]$$

Le TALM $\mathcal{T}^{\mathcal{A}}$ est associé à un sous ensemble de documents $D^{\mathcal{A}}$. Le caractère $\mathcal{A}$ désigne l'indicateur de contexte *sensible au temps*. $D^{\mathcal{A}}$ permet de constituer l'ensemble des documents utilisés pour les estimations de fréquence et de probabilité d'apparition des mots. Dans les équations suivantes, nous considérons l'ensemble de documents $D^{\mathcal{A}}$ où chaque document d est associé à une date t_d. Nous utilisons également t l'instant présent. La fonction $tf^{\mathcal{A}}(w, d)$ correspond à la fréquence d'apparition d'un mot w dans un document d, lissé selon $\Delta(t, t_d)$. Plus la date d'apparition d'un document t_d est éloignée de l'instant présent t, plus les fréquences d'apparition des mots sont diminuées. Nous définissons également $tf^{\mathcal{A}}(w, D^{\mathcal{A}})$ le nombre d'occurrences lissé d'un mot w dans l'ensemble des documents $D^{\mathcal{A}}$.

$$\begin{array}{rcl} tf^{\mathcal{A}}(w, d) & = & \Delta(t, t_d) \cdot tf(w, d) \\ tf^{\mathcal{A}}(w, D^{\mathcal{A}}) & = & \sum_{d \in D^{\mathcal{A}}} tf^{\mathcal{A}}(w, d) \end{array} \quad [5]$$

La fonction $len^{\mathcal{A}}(d)$ correspond à la somme des nombres d'occurrences de mots $w \in d$ lissés d'après la fonction $\Delta(t, t_d)$. La fonction $len^{\mathcal{A}}(w, D^{\mathcal{A}})$ correspond à la somme des tailles lissées dans l'ensemble des documents $D^{\mathcal{A}}$ considérés dans le modèle de langue $\mathcal{T}^{\mathcal{A}}$:

$$\begin{array}{rcl} len^{\mathcal{A}}(d) & = & \sum_{w \in d} tf^{\mathcal{A}}(w, d) \\ len^{\mathcal{A}}(D^{\mathcal{A}}) & = & \sum_{d \in D^{\mathcal{A}}} len^{\mathcal{A}}(d) \end{array} \quad [6]$$

À chaque document d peut être associé un poids qui dépend de $\Delta(t, t_i)$. Ainsi il est possible d'estimer un pseudo nombre de documents considérés dans le TALM en tenant compte de la distance temporelle qui sépare la date d'apparition du document et l'instant présent tel que :

$$N^{\mathcal{A}}(D^{\mathcal{A}}) = \sum_{d \in D^{\mathcal{A}}} \Delta(t, t_d) \quad [7]$$

Nous définissons également la fonction permettant le calcul de l'*inverse documents frequency* $idf^{\mathcal{A}}$ qui est souvent utilisé en recherche d'information comme indice du pouvoir discriminant d'un mot.

$$idf^{\mathcal{A}}(w, D^{\mathcal{A}}) = log \frac{N^{\mathcal{A}}(D^{\mathcal{A}})+1}{tf^{\mathcal{A}}(w, D^{\mathcal{A}})+0,5} \quad [8]$$

Soit $argmax(D^{\mathcal{A}})$ le dernier instant où le modèle $\mathcal{T}^{\mathcal{A}}$ a été mis à jour. Nous estimons alors la probabilité d'un mot w en fonction de l'instant présent t d'après les équations suivantes :

$$\begin{array}{rcl} p^{\mathcal{A}}(w|d) & = & \Delta(t, t_i) \cdot \frac{tf(w,d)}{len^{\mathcal{A}}(d)} \\ p^{\mathcal{A}}(w|D^{\mathcal{A}}) & = & \Delta(t, argmax(D^{\mathcal{A}})) \cdot \frac{\sum_{d \in D^{\mathcal{A}}} p^{\mathcal{A}}(w|d)}{N^{\mathcal{A}}(D^{\mathcal{A}})} \end{array} \quad [9]$$

4. Définition des méta critères pour le filtrage de documents

Nous définissons deux grandes familles de méta critères utiles pour déterminer si un document est centré sur une entité et s'il apporte ou non une information importante ou nouvelle. La première famille a pour but d'aider à la désambiguïsation de l'entité présente dans le document. La seconde famille de méta critères a pour but d'aider à déterminer si un document contient une information nouvelle ou importante.

4.1. *Les méta critères utiles à la désambiguïsation*

Afin d'estimer à quel point un document se réfère bien à l'entité cible plutôt qu'à l'un de ses homonymes, nous proposons de comparer la distribution lexicale de ce document avec celle du profil de l'entité en utilisant le RLM $\mathcal{R}$ à l'aide de la mesure de similarité cosine comme suit :

$$cos(d, \mathcal{R}) \quad = \quad \frac{\sum_{w \in d} tf(w,d).tf(w,\mathcal{R})}{\sqrt{\sum_{w \in d} tf(w,d)^2} \cdot \sqrt{\sum_{w \in d} tf(w,\mathcal{R})^2}} \qquad [10]$$

(Cucerzan, 2007) propose une approche qui permet d'identifier des relations entre l'entité recherchée et d'autres entités issues de Wikipedia. Si un document réfère à la fois à une entité, mais aussi à une de ses relations, alors il y a plus de chance que ce document se réfère réellement à l'entité que l'on recherche. Nous exploitons ces relations comme nouveaux critères et nous les catégorisons en trois types : entrante, sortante, réciproque (in, out et mut dans l'expression des critères).

Enfin, nous utiliserons un dernier critère qui donne une indication sur la manière dont le document parle de l'entité. Est-ce que le document est plutôt centré sur l'entité, ou est-ce que ce dernier la mentionne une seule fois ? En considérant l'ensemble des appellations V_e de l'entité cible e, nous pouvons estimer $p(V_e|d)$ la probabilité d'apparition de l'entité e (d'après toutes ses appellations) dans le document d d'après l'équation 11. Par exemple pour l'entité $e =$ *Tim_Cook*, on pourrait avoir un ensemble d'appellations comme $\{v_1 =$ *"Tim Cook"*$, v_2 =$ *"Le dirigeant d'Apple"*$, ...\}$. Nous utilisons toujours l'équation 11 pour estimer ce critère. Par ailleurs, le fait que l'entité soit mentionnée dans le titre peut également donner l'indication que le document est bien centré sur l'entité. L'équation 11 permet d'estimer par maximum de vraisemblance la probabilité d'apparition d'un mot s issu d'un ensemble de mots S en considérant un document d. Ainsi nous l'utilisons pour calculer les méta critères présentés dans le tableau 1.

$$p(S|d) = \frac{\sum_{s \in S} tf(s,d)}{|d|} \qquad [11]$$

4.2. *Méta critères basés sur l'analyse temporelle et le modèle de langue dynamique Time-Aware pour l'importance du document*

L'analyse de séries temporelles et particulièrement la détection du phénomène de rafale peuvent s'avérer utiles pour la détection ou même la prédiction d'évènements (Kleinberg, 2002 ; Sakaki *et al.*, 2010 ; Weng et Lee, 2011). Pour détecter si un document contient une information importante pour une entité, nous proposons

$p(V_e\|d_{title})$	probabilité d'apparition des variantes de nom V_e dans le titre
$p(V_e\|d)$	probabilité d'apparition des variantes de nom V_e dans le document
$p(Rel_{in}\|d)$	probabilité d'apparition des relations entrantes dans le document
$p(Rel_{out}\|d)$	probabilité d'apparition des relations sortantes dans le document
$p(Rel_{mut}\|d)$	probabilité d'apparition des relations mutuelles dans le document
$cos(d, \mathcal{R})$	cosinus entre un vecteur document d et le vecteur modèle de langue de référence $\mathcal{R}$

Tableau 1. *Méta critères permettant la désambiguïsation de l'entité dans un document en utilisant les variantes de nom et les relations entre entités.*

d'utiliser des méta critères qui caractérisent à la fois la nouveauté et le phénomène de rafale. Bien sûr ce phénomène de rafale peut être plus ou moins vérifié selon la célébrité de l'entité cible, les lignes éditoriales des auteurs des documents du flux observé et les autres actualités du moment (une actualité très dense tend à diminuer l'effet rafale puisque les journaux sont de taille limitée et qu'ils ne peuvent alors se permettre trop de redondance).

Une entité (et donc son profil) évolue au fur et à mesure que le temps passe. Il a été montré dans plusieurs études (Wang *et al.*, 2007 ; Amodeo *et al.*, 2011 ; Peetz *et al.*, 2014) que l'analyse de pertinence d'un document peut être améliorée en considérant des phénomènes de rafale sur les requêtes émises ou sur les documents nouveaux. La figure 1 montre un exemple de phénomène de rafale pour l'entité *BNSF railway*. Cette rafale a lieu peu de temps après qu'un accident se soit produit sur ces lignes de chemin de fer. Nous choisissons d'utiliser l'algorithme de (Kleinberg, 2002) pour mesurer la force de la rafale à chaque instant. Un autre critère est également proposé par (Diaz et Jones, 2004). Ils proposent d'utiliser le coefficient d'aplatissement (Kurtosis) d'une série temporelle X où chaque élément X_i correspond au nombre de documents qui apparaissent chaque unité temporelle. Si l'unité correspond à *1 heure*, alors chaque élément $X_i \in X$ aurait pour valeur nombre de documents qui mentionnent l'entité à partir d'un instant t jusqu'à l'instant $t + 1h$.

Nous avons formalisé dans la section 3.2 un modèle de langue sensible à l'évolution temporelle (Time-Aware Language Model TALM). Le TALM a pour vocation d'être mis à jour avec les documents qui apparaissent sur un flux. Le TALM a conscience du temps lors des mises à jour ainsi, il peut estimer la probabilité d'apparition des mots à un instant t. Pour mesurer le degré de nouveauté des documents qui apparaissent sur le flux, nous proposons d'utiliser une mesure de divergence entre le TALM et le modèle de référence RLM. Le TALM étant le reflet de ce qui se passe autour de l'entité à l'instant t, l'idée est de le comparer avec ce qui est déjà connu à propos de l'entité. Nous proposons d'utiliser la divergence de Jensen-Shannon (JSD), qui est symétrique. La JSD s'utilise avec deux vecteurs de probabilités A et B. La symétrie est possible en calculant un vecteur C résultant de la moyenne entre A et B (équation 12).

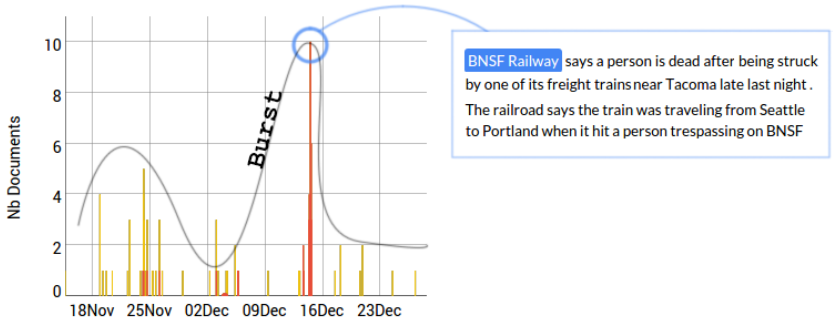

Figure 1. *Phénomène de rafale sur une information vitale pour l'entité BNSF Railway, lors d'accident sur une ligne de chemin de fer.*

Notre représentation du RLM et du TALM permettent de déduire deux vecteurs de probabilités à partir des formules données dans les équations 2 et 9. Il est également possible d'utiliser l'équation 2 pour estimer un vecteur de probabilités pour un document d. Nous pouvons alors utiliser cette mesure de divergence comme méta critère en utilisant différentes combinaisons (cf., tableau 2).

$$\begin{aligned} C &= \tfrac{1}{2} \times (A + B) \\ JSD(A,B) &= \tfrac{1}{2} \times \textstyle\sum_{w \in C} p(w|A) log \frac{p(w|A)}{p(w|C)} \\ &\quad + \tfrac{1}{2} \times \textstyle\sum_{w \in C} p(w|B) log \frac{p(w|B)}{p(w|C)} \end{aligned} \quad [12]$$

(Karkali *et al.*, 2014) ont testé différentes approches pour mesurer la nouveauté sur des données réelles. Le score de nouveauté qui offre les meilleurs résultats est celui qui utilise une version lissée du $tf.idf$ à laquelle est ajoutée une composante temporelle. Nous considérons $V_t^{\mathcal{A}}$ une représentation d'un vecteur de probabilité du TALM à un instant t. Le score de nouveauté $NS^{\mathcal{A}}(A, V_t^{\mathcal{A}})$ pour l'instant t est formulé dans l'équation 13.

$$NS^{\mathcal{A}}(A, V^{\mathcal{A}}) = \frac{1}{\sum_{w \in A} tf(w,A)} \times \textstyle\sum_{w \in A} tf(w,A) \cdot idf^{\mathcal{A}}(w, V_t^{\mathcal{A}}) \quad [13]$$

(Carbonell et Goldstein, 1998) ont présenté la mesure *Maximal Marginal Relevance* (MMR) qui permet la combinaison d'un score de pertinence d'après une requête et d'un score de nouveauté pour un contexte de classement de documents. Nous nous sommes inspiré de la formulation présentée pour formuler un score de MMR. Nous proposons d'utiliser à la fois le score issu du méta critère $cos(d, \mathcal{R})$ et $JSD(d, V_t^{\mathcal{A}})$:

$$MMR(d, \mathcal{R}, \mathcal{T}^{\mathcal{A}}) = \alpha.cos(d, \mathcal{R}) - (1 - \alpha).JSD(d, V_t^{\mathcal{A}}) \quad [14]$$

$Kleinberg(X_e)$	Force de la rafale de la série temporelle X_e de l'entité e.
$Kurtosis(X_e)$	Coefficient d'aplatissement de la série temporelle X_e de l'entité e.
$JSD(\mathcal{R}, \mathcal{T}^{\mathcal{A}})$	Divergence de Jensen Shannon entre le RLM $\mathcal{R}$ et le TALM $\mathcal{T}^{\mathcal{A}}$.
$JSD(d, \mathcal{T}^{\mathcal{A}})$	Divergence de Jensen Shannon entre un nouveau document d et le TALM $\mathcal{T}^{\mathcal{A}}$.
$NS(\mathcal{R}, \mathcal{T}^{\mathcal{A}})$	Score de nouveauté entre le RLM $\mathcal{R}$ et le TALM $\mathcal{T}^{\mathcal{A}}$ en considérant l'instant t_c.
$NS(d, \mathcal{T}^{\mathcal{A}})$	Score de nouveauté entre un nouveau document d et le TALM $\mathcal{T}^{\mathcal{A}}$ en considérant l'instant t_c.
$MMR(d, \mathcal{R}, \mathcal{T}^{\mathcal{A}})$	Combinaison des scores de pertinence et de nouveauté pour un nouveau document d à l'aide du RLM $\mathcal{R}$ et du TALM $\mathcal{T}^{\mathcal{A}}$ en considérant l'instant t_c.

Tableau 2. *Méta critères qui permettent de qualifier l'intérêt d'un document.*

Dans le tableau 2, nous résumons les différents méta critères utiles pour qualifier l'importance d'un document en considérant des aspects de nouveauté, de diversité et de rafale.

5. Expérimentations

5.1. *Description de la tâche Knowledge Base Acceleration (KBA)*

Les pistes KBA des évaluations TREC sont directement liées au problème de maintien à jour de bases de connaissances. En effet, les bases de connaissances, comme Wikipédia, sont difficiles à maintenir à jour de par le nombre immense d'articles en rapport au nombre de contributeurs actifs régulièrement. Le scénario de KBA est de simplifier la mise à jour les bases de connaissances de manière automatique en suggérant les articles intéressants pour un sujet en particulier.

Pour permettre une simulation à grande échelle, les organisateurs de la tâche ont mis en place un corpus de documents datés (d'octobre 2011 à mai 2013) simulant un flux de documents. Il contient plus d'un milliard de documents issus du Web et plus spécifiquement de sites d'actualité, de forums, de blogs. Le corpus peut être parcouru de manière chronologique. Les organisateurs de la tâche ont sélectionné une centaine d'entités et ont annoté manuellement des documents selon 4 classes :

- Garbage : le document ne concerne pas du tout l'entité ;

- Neutral : le document mentionne l'entité, mais n'est pas centré sur elle ;

- Useful : le document est centré sur l'entité et n'apporte pas d'information nouvelle ;

- Vital : le document est centré sur l'entité et apporte une information nouvelle.

Chaque année, les organisateurs sélectionnent un certain nombre d'entités à partir de critères qui rendent la tâche encore plus complexe. Les entités ne sont pas ou peu populaires et elles peuvent être homonymes. Les participants doivent, pour chacune des entités, filtrer les documents du flux et leur attribuer une classe parmi les 4 classes : *garbage, neutral, useful* et *vital*. La décision d'attribuer une classe à un document doit se faire dès le moment où un document est évalué. Bien entendu, le corpus ne peut être indexé et ce dernier doit être parcouru en considérant l'ordre chronologique. Nous utiliserons pour cette évaluation les données de KBA 2013. Certaines entités sont fournies avec un à trois documents de références afin d'initialiser le système. Ces documents sont utilisés pour initialiser le RLM de chacune des entités. Pour certaines entités, aucun document de référence n'est proposé, dans ce cas nous utilisons les deux premiers documents dits utiles qui se trouvent dans les données d'entraînement.

5.2. *Les différentes stratégies d'expérimentation*

Notre système fait un filtrage en deux étapes. La première vérifie que le document contient bien une mention de l'entité à l'aide des formes de surface. Pour la seconde étape, nous utilisons un système de classification de type forêts aléatoires *Random Forest* que l'on entraîne en calculant les valeurs des méta critères présentés en section 4 pour les documents annotés manuellement. L'entraînement conduit à hiérarchiser les méta critères et estimer, pour chacun d'eux, les valeurs limites correspondant à la meilleure prise de décision. Une fois la forêt apprise sur les entités d'entraînement, elle peut être appliquée en test sur n'importe quelle nouvelle entité sans réentraînement.

Pour vérifier l'impact de l'utilisation du modèle de langue *Time-Aware* (TALM), nous proposons différentes stratégies. La première stratégie, *NU*, n'utilise pas du tout le TALM, et donc les critères qui lui sont associés ne sont pas pris en compte. La seconde stratégie, *US*, met à jour le TALM en n'utilisant qu'un extrait du document (le snippet) . Cet extrait est construit en concaténant les paragraphes qui contiennent au moins une mention de l'entité. La dernière stratégie, *UD*, met à jour le TALM en utilisant le document complet. Nous avons également mis en place différents systèmes pour la classification. Le premier système, *2STEPS*, considère le problème comme étant un problème de classification binaire. Deux classifieurs sont utilisés en cascade. Le premier fait une classification selon les deux classes *Garbage/Neutral* et *Useful/Vital*. Le second classifieur ne classe que les documents classés *Useful/Vital* par le premier. Il tente alors de déterminer si le document est plutôt *useful* ou *vital*. Le second système, *SINGLE*, fait une classification directement sur les 4 classes. Le troisième système, *VvsAll*, essaye de déterminer la classe d'un document parmi les deux classes *Garbage/Neutral/Useful* et *Vital*. Dans le cas où le document n'est pas *vital*, ce dernier est soumis à un classifieur qui détermine la classe entre *Garbage, Neutral* ou *Useful*. Le dernier système, *MULTI*, fait la synthèse des scores donnés par les trois précédentes stratégies pour déterminer une classe parmi *Garbage, Neutral, Useful* ou *Vital*. Enfin, pour l'estimation des paramètres de la fonction temporelle,

nous avons défini de manière arbitraire la valeur suivante $\rho = 10$. Pour le paramètre lambda nous avons utilisé la validation croisée et nous avons déterminé que le meilleur résultat était obtenu avec $\lambda = 14jours$ cette valeur étant convertie en seconde pour l'expérimentation.

5.3. *Analyse des résultats*

Pour observer les performances du système, nous avons utilisé l'outil d'évaluation officiel. La mesure utilisée pour l'étude est la F-mesure $f_1(p, r))$ qui est la moyenne harmonique de la précision et du rappel. Les scores de KBA sont observés en utilisant plusieurs seuils suivant le score de confiance donné par l'algorithme de classification (score allant de 0 à 1000). Dans le tableau 3, nous pouvons voir les scores que nous obtenons pour la classification de documents appartenant à la classe *vital* seulement. Certains systèmes offrent des performances supérieures au meilleur résultat présenté lors de la campagne d'évaluation de 2013. Aussi, les meilleurs résultats sont obtenus sur le système MULTI pour chacune des stratégies de mise à jour du TALM. Cela montre que les systèmes de classification peuvent être complémentaires. En règle générale, l'utilisation du modèle sensible au temps tend à faire baisser les performances pour cette classe.

Systèmes	**F-mesure - KBA 2013**		
	NU	**UD**	**US**
MULTI	**0,435**	0,356	0,381
SINGLE	**0,389**	0,311	0,358
2STEPS	**0,341**	0,213	0,222
VvsAll	0,250	**0,288**	0,284
Score officiel médian KBA 2013	0,201		
Meilleur système KBA 2013	0,360		

Tableau 3. *Scores issus du logiciel d'évaluation officiel TREC KBA pour la classification de documents de classe "Vital" pour les différents systèmes et stratégies de mise à jour que nous proposons.*

Nous avons voulu en savoir plus sur les critères prédominants dans le processus de décision de la classe *vital*. Pour cela, nous avons utilisé le logiciel R et la bibliothèque « Party » qui implémente un algorithme de classification de type *« Random-Forest »* pour lequel il est possible de calculer les Variables d'Importances (VI)(Breiman, 2001). La figure 2 montre les variables classées de la plus importante (en haut) à la moins importante (en bas). Nous avons tracé les variables d'importance calculées sur la partie entraînement et la partie test. En ce qui concerne la partie entraînement, nous remarquons que les méta critères qui utilisent des analyses temporelles (c.-à-d., Kleinberg, Kurtosis) sont moins déterminants pour la prise de décision finale. Ce résultat est surprenant dans le sens où cela montre qu'une rafale de documents n'est pas forcément un indicateur décisif pour la détection d'évènements centrés sur une

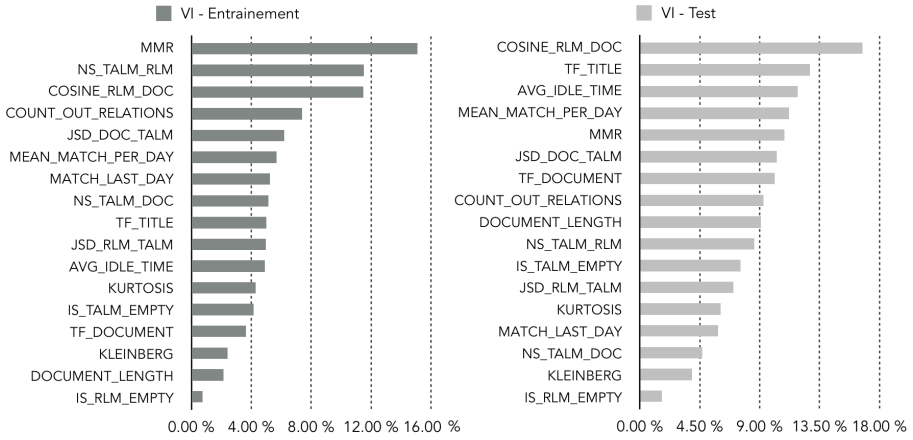

Figure 2. *Variables d'importances calculées sur les données d'entraînement et de test à l'aide du système 2Steps-US (pour le classifieur Useful vs Vital) avec la mise à jour du TALM en utilisant les snippets.*

entité. En revanche, on remarque que les méta critères relatifs à la désambiguïsation et à la nouveauté sont parmi les plus importants. Par ailleurs, le critère qui utilise les entités en relation avec l'entité recherchée est également très important dans la prise de décision (`COUNT_OUT_RELATIONS`). On remarquera également qu'il y a beaucoup de changement entre l'entraînement et le test ce qui veut dire que les données d'entraînement ne sont pas assez représentatives de ce qu'il y a dans le test.

L'enjeu de la tâche KBA est de détecter automatiquement les documents *vitaux*. La classe *Useful* reste néanmoins une classe importante. Détecter de manière précise qu'un document est *useful* permet de filtrer déjà les informations qui concernent les entités recherchées. Le tableau 4 montre les résultats obtenus pour les classes *Useful* et *Vital* avec le meilleur système (MULTI). Nous remarquons que nous obtenons des scores proches, voire supérieurs, au meilleur système de KBA.

Systèmes	**F-mesure - KBA 2013**		
	NU	**UD**	**US**
MULTI	**0,715**	0,637	0,639
Score officiel médian KBA 2013	0,406		
Meilleur système KBA 2013	0,659		

Tableau 4. *Scores issus du logiciel d'évaluation TREC KBA officiel pour la classification de documents de classe Vital/Useful pour les différents systèmes et stratégies de mise à jour que nous avons proposés (données KBA 2013).*

6. Conclusion

Dans cet article nous avons présenté les profils d'entités basées sur deux modèles de langue. Le premier retranscrit les connaissances à caractère biographique de l'entité, le second retranscrit l'actualité de l'entité. Ainsi la partie du profil qui permet la désambiguïsation n'est jamais altérée. Pour le second modèle, nous avons formalisé un modèle de langue sensible au temps afin que ce dernier oublie les informations les plus anciennes. Nous évitons ainsi la dérive du modèle. Grâce au fait que l'oubli se fasse progressivement, il est toujours possible de détecter la nouveauté dans l'actualité de l'entité. Nous proposons finalement un ensemble de méta critères basés sur ces nouveaux profils d'entités. Nous les utilisons dans un système de classification pour filtrer les documents vitaux issus d'un flux de documents. Nous avons évalué notre approche dans l'environnement de la tâche Knowledge Base Acceleration (KBA) de la conférence TREC (Text REtrieval Conference). Nous avons montré que nos systèmes offrent des performances supérieures à celles présentées jusqu'à présent.

En perspective, nous souhaiterions améliorer notre approche tout en restant le moins supervisé possible. Nous travaillons actuellement sur un système de regroupement par types d'entités afin de calculer des modèles de classification plus ciblés et plus performants.

7. Bibliographie

Amodeo G., Amati G., Gambosi G., « On relevance, time and query expansion », *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, p. 1973-1976, 2011.

Balog K., Ramampiaro H., Takhirov N., Nørvåg K., « Multi-step classification approaches to cumulative citation recommendation », *Open research Areas in Information Retrieval, OAIR '13, Lisbon, Portugal, May 15-17, 2013*, p. 121-128, 2013.

Belkin N. J., Croft W. B., « Information filtering and information retrieval : two sides of the same coin ? », *Communications of the ACM*, vol. 35, n° 12, p. 29-38, December, 1992.

Bellogín A., Gebremeskel G., « CWI and TU Delft Notebook TREC 2013 : Contextual Suggestion, Federated Web Search, KBA, and Web Tracks », *proceedings of the Twenty-Second TREC 2013*, National Institute of Standards and Technology (NIST), p. 500-302, 2014.

Bonnefoy L., Bouvier V., Bellot P., « LSIS/LIA at TREC 2012 knowledge base acceleration », *Proceedings of the Twenty-First TREC 2012*, p. 500-298, 2013a.

Bonnefoy L., Bouvier V., Bellot P., « A weakly-supervised detection of entity central documents in a stream », *The 36th International ACM SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013*, ACM, p. 769-772, 2013b.

Breiman L., « Random Forests », *Machine Learning*, vol. 45, n° 1, p. 5-32, 2001.

Carbonell J. G., Goldstein J., « The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries », *SIGIR '98 : Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, ACM, p. 335-336, 1998.

Cucerzan S., « Large-Scale Named Entity Disambiguation Based on Wikipedia Data », *Proceedings of the 2007 Joint Conference EMNLP-CoNLL 2007*, p. 708-716, 2007.

Diaz F., Jones R., « Using temporal profiles of queries for precision prediction », *in* M. Sanderson, K. Järvelin, J. Allan, P. Bruza (eds), *SIGIR 2004 : Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, ACM, p. 18-24, 2004.

Efron M., « The University of Illinois' Graduate School of Library and Information Science at TREC 2013 », *proceedings of the Twenty-Second TREC 2013*, National Institute of Standards and Technology (NIST), 2014.

Endres D. M., Schindelin J. E., « A new metric for probability distributions », *IEEE Transactions on Information Theory*, vol. 49, n° 7, p. 1858-1860, 2003.

Frank J., Kleiman-Weiner M., Roberts D. A., Niu F., Zhang C., « Building an Entity-Centric Stream Filtering Test Collection for TREC 2012 », *Proceedings of the Twenty-First TREC 2012*, National Institute of Standards and Technology (NIST), p. 500-298, 2012.

Karkali M., Rousseau F., Ntoulas A., Vazirgiannis M., « Using temporal IDF for efficient novelty detection in text streams », *CoRR*, 2014.

Kjersten B., McNamee P., « The HLTCOE approach to the TREC 2012 KBA track », *Proceedings of The 21th TREC*, 2012.

Kleinberg J., « Bursty and hierarchical structure in streams », *Proceedings of the Eighth ACM SIGKDD 2002, Edmonton, Alberta, Canada*, p. 91-101, 2002.

Li X., Croft W. B., « Time-based language models », *Proceedings of the 2003 ACM CIKM 2003*, p. 469-475, 2003.

Liu X., Fang H., « Leveraging related entities for knowledge base acceleration », *in* Y. Zeng, S. Kotoulas, Z. Huang (eds), *Proceedings of the 4th international workshop on Web-scale knowledge representation retrieval and reasoning, Web-KR@CIKM 2013, San Francisco, CA, USA, November 1, 2013*, ACM, p. 1-4, 2013.

Navigli R., « Word sense disambiguation : A survey », *ACM Comput. Surv.*, 2009.

Peetz M.-H., Meij E., de Rijke M., « Using temporal bursts for query modeling », *Inf. Retr.*, vol. 17, n° 1, p. 74-108, 2014.

Sakaki T., Okazaki M., Matsuo Y., « Earthquake Shakes Twitter Users : Real-time Event Detection by Social Sensors », *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, ACM, New York, NY, USA, p. 851-860, 2010.

Sehgal A. K., Srinivasan P., « Profiling Topics on the Web », *Proceedings of the WWW2007 Workshop*, p. 1-8, 2007.

Wang X., Zhai C., Hu X., Sproat R., « Mining correlated bursty topic patterns from coordinated text streams », *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*, p. 784-793, 2007.

Weng J., Lee B., « Event Detection in Twitter », *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.