Modèles de langue pour la mise à jour d'un profil d'entité

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

Institut de Recherche en Informatique de Toulouse, SIG
118, route de Narbonne, F-31062 Toulouse Cedex 9, France
{rafik.abbes, karen.sauvagnat, nathalie.hernandez, mohand.boughanem}@irit.fr

RÉSUMÉ. *Dans cet article nous souhaitons renvoyer à partir de documents issus du Web, ceux apportant des informations nouvelles sur une entité donnée. Ces documents peuvent ainsi servir à mettre à jour un profil existant (par exemple une page Wikipedia) de cette entité. Notre approche se base initialement sur un appariement des mentions de l'entité afin de renvoyer un premier ensemble de documents pertinents, puis s'appuie sur des modèles de langue estimés à partir de différentes unités d'information. Nous avons évalué notre approche dans le cadre de la tâche "Cumulative Citation Recommendation" de TREC KBA* 2013*. Les résultats montrent l'intérêt des modèles de langue par rapport aux méthodes de l'état de l'art, et que la vitalité est mieux estimée en considérant tout le contenu des documents mentionnant l'entité.*

ABSTRACT. *In this paper, we aim at identifying vital documents that a human would want to cite when updating an entity profile (for example, its Wikipedia article). In our approach, we first retrieve a set of potential relevant documents containing at least one entity mention, then we rank vital documents using vitality-based language models estimated from different information units. We evaluated our approach through the 2013 CCR task of TREC KBA. Results show the interest of our approach compared to those of the state-of-the-art. We also show that vitality is better estimated when considering the whole content of documents mentioning the entity instead of considering only the entity sentences.*

MOTS-CLÉS : *requêtes entités, profil d'entité, modèle de langue de vitalité*

KEYWORDS: *entity queries, entity profile, vitality-based language model*

CORIA 2014, pp. 137–151, Nancy, 18-21 mars 2014

1 Introduction

Selon une étude récente (Pound *et al.,* 2010), plus de la moitié des requêtes du Web ciblent une ou plusieurs entités. Une entité est un objet, une chose concrète ou abstraite qui peut être distinctement identifiée, par exemple une organisation, une personne, un événement, une date ayant connu des événements particulier, etc..

Aujourd'hui, Wikipedia est une des principales sources consultées par les utilisateurs pour accéder aux connaissances disponibles sur une grande variété d'entités (Li *et al.,* 2012). Quatre millions d'articles relatifs à des entités sont disponibles en 2013, selon les statistiques actuelles du site Wikipedia[1].

Toutefois, en raison du grand nombre d'articles Wikipedia, de nombreux contenus ne sont pas examinés par des experts, de sorte que le nombre de textes de mauvaise qualité a également augmenté considérablement (Suzuki *et al.,* 2013). De plus, à cause du nombre limité de contributeurs (15 000) par rapport au nombre d'entités (4 398 244), reporter une nouvelle information en rapport avec une entité donnée sur sa page Wikipedia se fait généralement avec un temps de latence médian de 356 jours (Frank *et al.,* 2012). Ce problème de temps de latence a été évoqué récemment dans la tâche *Knwoledge Base Acceleration* (KBA) de TREC[2]. Pour résoudre ce problème, une sous-tâche de TREC KBA a été proposée, nommée *Cumulative Citation Recommendation* (CCR). Cette tâche consiste à recommander aux contributeurs d'un profil d'une entité, des documents présents sur le Web qui contiennent de l'information pertinente sur cette entité. Le profil d'une entité donnée est une page qui décrit cette entité, par exemple une page Wikipedia ou Freebase.

Deux scénarios possibles peuvent être identifiés :

– **Scénario 1 :** Un contributeur veut créer un nouveau profil sur une entité donnée. Dans ce cas, tous les documents qui parlent de cette entité pourraient l'aider. Ces documents sont appelés documents *centrés* sur l'entité.

– **Scénario 2 :** Un contributeur veut mettre à jour un profil existant sur une entité donnée. Dans ce cas, seuls les documents ajoutant une information nouvelle sur cette entité pourraient lui servir. Ces documents sont appelés documents *vitaux*. Les autres documents centrés sur l'entité mais répétant des informations déjà données dans le profil de l'entité sont *redondants*.

Durant la première année de la tâche CCR (en 2012), le but était d'identifier à partir d'un flux de documents issus du Web, ceux qui sont centrés sur une entité donnée, sans distinguer entre les documents redondants et les documents vitaux. Dans ce contexte, un contributeur voulant mettre à jour un profil existant (scénario 2) pouvait se trouver submergé par plusieurs documents centrés sur l'entité, mais n'apportant aucune information nouvelle par rapport à ce profil. En 2013, la tâche de CCR a évolué,

1. http ://en.wikipedia.org/wiki/Wikipedia :Size_of_Wikipedia (Décembre 2013)
2. http ://trec-kba.org/

et les systèmes devaient distinguer entre les documents redondants et les documents vitaux, afin de mettre à jour un profil d'entité existant.

Une variété d'approches a été proposée se basant sur l'appariement simple du nom de l'entité (Liu *et al.,* 2013), l'expansion de la requête par des noms d'entités reliées (Dietz *et al.,* 2013) (Zhang *et al.,* 2013) ou l'utilisation des classifieurs (Bonnefoy *et al.,* 2013) (Wang *et al.,* 2013) (Abbes *et al.,* 2013).

Ces approches ont montré qu'elles pouvaient atteindre un bon taux de rappel pour les documents centrés sur une entité donnée, mais elles ne permettent pas de bien distinguer les documents vitaux des documents redondants.

Dans cet article, nous proposons d'exploiter les modèles de langue pour modéliser la notion de vitalité. Nous souhaitons en particulier répondre aux questions suivantes :

– Quelles sont les informations à considérer pour estimer un modèle de langue permettant de générer des documents vitaux ? Tout le contenu d'un document vital est-il nécessaire à cette estimation ?

– Comment estimer le modèle de langue à partir de ces informations ?

Cet article est organisé comme suit. Nous présentons dans la section 2 un état de l'art se focalisant sur la recherche de documents centrés sur une entité. La section 3 présente l'approche proposée pour l'identification des documents vitaux. Dans la section 4, nous présentons et discutons les résultats de notre approche et nous les comparons aux systèmes participants de la tâche TREC KBA 2013. Nous concluons et énonçons quelques perspectives en section 5.

2 État de l'art

2.1 Recherche de documents centrés sur une entité

L'émergence des bases de connaissances comme Wikipedia et Freebase a engendré de nouvelles problématiques dans le domaine de la recherche d'information, comme l'interprétation et l'expansion de la requête via une base de connaissance (Aggarwal *et al.,* 2012) (Pound *et al.,* 2012), la recherche d'entités répondant à un besoin en information (Bron *et al.,* 2010), et la recherche de documents centrés sur une entité.

Nous nous intéressons dans ce travail à cette dernière problématique où nous devons faire face à deux difficultés : la première est la *désambiguïsation* des mentions d'entités dans un document : une même mention d'entité peut faire référence à plus d'une entité, par exemple la mention 'James Parsons' peut faire référence à l'avocat 'James A. Parson', au juge 'James B. Parsons', ou l'acteur 'James Joseph Parsons'. La deuxième difficulté est la *variété* de mentions pour une même entité. Par exemple, l'acteur 'James Joseph Parsons' peut être référencé par plusieurs mentions possibles (variantes) comme : 'James J. Parson', 'James Parson', 'Jim Parsons', etc.

Pour rechercher les documents centrés sur une entité, une variété d'approches a été proposée. Ces approches ont été évaluées dans le cadre des tâches *Cumulative Citation Recommendation* de TREC KBA 2012 et 2013 (Frank *et al.,* 2012) (Frank *et al.,* 2013). La tâche de 2013 se focalisant particulièrement sur la recherche de documents vitaux pour un profil d'entité, nous nous focaliserons sur les approches proposées dans ce cadre. La méthode proposée dans (Liu *et al.,* 2013) calcule un score pour chaque document, en prenant en compte l'existence d'au moins une mention d'entité cherchée et aussi du nombre et du poids des entités reliées mentionnées. D'autres méthodes ont utilisé la technique d'expansion de la requête par des entités reliées (Dietz *et al.,* 2013) (Zhang *et al.,* 2013). Ces méthodes permettent d'avoir un bon taux de rappel pour les documents centrés sur une entité, mais elles n'arrivent pas à bien distinguer les documents vitaux des documents redondants. D'autres méthodes se basent sur une classification supervisée (Wang *et al.,* 2013) (Abbes *et al.,* 2013). Par exemple dans (Wang *et al.,* 2013), les auteurs ont utilisé des forêts d'arbres de décision afin d'identifier la classe (vital, redondant, non pertinent) d'un document donné *d* en utilisant cinq familles de critères : des critères liés au document *d* comme la longueur et la source, des critères liés à l'entité comme le nombre d'entités reliées, des critères décrivant la relation entre le document *d* et l'entité comme le nombre de mentions de l'entité dans le document, des critères temporels, et des critères de citation calculant la similarité entre le document *d* et des documents déjà cités dans le profil de l'entité. Cette méthode distingue mieux les documents vitaux des document redondants, mais les critères de citation qu'elle utilise ne peuvent pas être appliqués pour les profils d'entités ne citant au départ aucun document.

Dans cet article, nous proposons une méthode ayant pour but d'estimer la vitalité d'un document pertinent par rapport à une entité donnée en nous inspirant de quelques travaux traitant de la nouveauté dans les textes.

2.2 Recherche de la nouveauté dans les textes

L'aspect de nouveauté dans les textes a été étudié dans divers travaux de recherche d'information et traité dans des compagnes d'évaluation comme TREC (TREC Novelty) (Harman, 2002). Le but était d'éviter de restituer les informations redondantes par rapport à des informations déjà consultées par l'utilisateur. La méthode proposée dans (Larkey *et al.,* 2002) calcule une mesure de nouveauté en se basant sur le comptage simple des mots nouveaux n'apparaissant pas dans les documents (ou phrases) pertinents déjà restitués à l'utilisateur. Dans (Zhang *et al.,* 2002), d'autres mesures ont été proposées, en se basant sur le calcul de la distance de cosinus entre le document en cours (d_t) et les documents déjà consultés par l'utilisateur (DT), ou sur le calcul de la divergence entre le modèle de langue de d_t et le modèle de langue de DT.

Dans cet article, nous calculons une mesure de vitalité similaire à la mesure basée sur les modèles de langue proposée par (Zhang *et al.,* 2002), mais dans un contexte différent : nous ne nous intéressons pas à la nouveauté d'un document résultat par rapport aux documents déjà recommandés à un contributeur, mais plutôt aux docu-

ments apportant une information nouvelle par rapport à un profil d'entité existant. Nous calculons ainsi cette mesure de vitalité en tenant compte du modèle de langue du document en cours et d'un modèle de langue de vitalité.

3 Une approche basée sur le modèle de langue pour l'identification des documents vitaux

Notre but est d'identifier à partir d'un ensemble de documents centrés sur une entité donnée, ceux qui sont vitaux, c'est à dire ceux qui permettraient de mettre à jour son profil (par exemple sa page Wikipedia). Intuitivement, nous supposons qu'un document vital utilise un ensemble de termes qui peuvent refléter la vitalité. Par exemple, un document citant cette phrase *"Celine Dion vient de dévoiler son nouvel album "Loved Me Back to Life""*, a de fortes chances d'apporter une information nouvelle par rapport à un profil existant de l'entité *"Celine Dion"*, puisqu'il utilise des termes pouvant refléter la vitalité comme *'vient'*, *'dévoiler'*, *'nouvel'*.

Dans notre approche, nous proposons de modéliser la notion de vitalité d'un document en nous basant sur les modèles de langue (Ponte *et al.,* 1998). Nous proposons d'estimer un modèle de langue nommé modèle de vitalité, générant des documents vitaux par rapport à une entité donnée. Pour estimer le modèle vital d'une entité, nous supposons disposer d'un échantillon de documents vitaux pour cette entité. Nous notons cet ensemble vital $EV_e = \{dv_1, dv_2, \ldots, dv_m\}$.

dv_i représente tout le contenu ou une partie d'un document échantillon vital. Une seule partie de document (phrase, paragraphe, etc.) pourrait en effet être suffisante pour estimer sa vitalité. Dans la suite de l'article, par abus de langage, nous parlerons de document échantillon vital dv_i, dv_i représentant cependant tout ou une partie d'un document échantillon vital.

La vitalité peut être estimée de deux façons :

1) Elle peut être considérée comme unidimensionnelle. Nous pouvons alors estimer un seul modèle de vitalité θ_{Ve_u} à partir d'un seul document DV_e qui représente la concaténation de tous les documents échantillons vitaux de l'ensemble EV_e.

2) Elle peut être considérée comme multidimensionnelle. Nous pouvons dans ce cas estimer un sous-modèle de vitalité θ_{dv_i} pour chaque dv_i de l'ensemble EV_e.

3.1 Estimation d'un modèle de vitalité unidimensionnel

Étant donné un ensemble de documents échantillons vitaux EV_e pour une entité donnée, la probabilité de générer un terme t à partir d'un modèle de vitalité unidimensionnel θ_{Ve_u} est estimée de la façon suivante (lissage de Dirichlet) :

$$P(t|\theta_{Ve_u}) = \frac{tf(t, DV_e) + \mu\ P(t|C)}{|DV_e| + \mu} \quad [1]$$

où

DV_e représente un document concaténant tous les documents échantillons vitaux appartenant à l'ensemble EV_e

C est une collection de référence comportant tous les documents d'apprentissage et tous les documents de test.

$tf(t, DV_e)$ représente la fréquence d'apparition du terme t dans le document DV_e

μ représente une valeur de lissage réelle $\in [0, +\infty[$

$P(t|C) = \frac{tf(t,C)}{\sum_{t' \in T} tf(t',C)}$, avec T représente tous les termes du vocabulaire.

3.2 Estimation d'un modèle de vitalité multidimensionnel

Étant donné un ensemble de documents échantillons vitaux EV_e correspondant à une entité e composée de n termes, en faisant l'analogie par rapport au modèle de pertinence (Lavrenko *et al.,* 2001), la probabilité de générer un terme t à partir d'un modèle de vitalité multidimensionnel θ_{Ve_m} est estimée comme suit :

$$\begin{aligned} P(t|\theta_{Ve_m}) &= \sum_{i=1}^{m} P(t|\theta_{dv_i}) P(\theta_{dv_i}|e) \\ &\propto \sum_{i=1}^{m} P(t|\theta_{dv_i}) P(e|\theta_{dv_i}) \end{aligned} \quad [2]$$

$$P(e|\theta_{dv_i}) = \prod_{t \in e} P(t|\theta_{dv_i}) \quad [3]$$

où

$P(t|\theta_{dv_i})$ est l'estimation d'un modèle de vitalité unidimensionnel à partir d'un seul document vital dv_i. Cette probabilité est calculée selon l'équation 1.

m représente le nombre de documents jugés vitaux pour l'entité e.

3.3 Mesure de la vitalité d'un document

Soit un nouveau document d composé de n termes t.

Le score de vitalité du document d par rapport à un modèle de vitalité d'une entité θ_{Ve_x} (unidimensionnel θ_{Ve_u} ou multidimensionnel θ_{Ve_m}) est traduit par la vraisemblance des termes du modèle vital :

$$Score_{Vitalité1}(d, e) = \prod_{t \in Ve_x} P(t|\theta_d)^{P(t|Ve_x)} \quad [4]$$

où $P(t|\theta_d)$ est calculée selon l'équation 1.

4 Expérimentation et résultats

Nous avons évalué notre méthode dans le cadre de la tâche CCR[3] de TREC KBA 2013. Dans cette section, nous présentons une description générale de cette tâche, ensuite nous comparons les différentes configurations de notre approche. Enfin, nous positionnons notre approche par rapport aux autres méthodes proposées dans la tâche.

4.1 Description de la tâche TREC KBA CCR 2013

La tâche CCR de TREC KBA 2013 consiste à analyser un flux de documents issus du Web afin d'identifier ceux qui sont vitaux par rapport à un profil d'entité donné (par exemple sa page Wikipedia).

Le corpus[4] élaboré dans le cadre de cette tâche comporte plus de 500 millions de documents issus de plusieurs sources (Presse, Web, Social, Forum, Blog, etc.) ayant une taille de 4.5 Téra octets compressés. Les documents sont datés dans la période allant du mois d'octobre 2011 jusqu'au mois de février 2013.

Le corpus est divisé en deux périodes :

– une période d'apprentissage allant d'octobre 2011 à février 2012, durant laquelle chaque document mentionnant l'entité (requête) a été annoté manuellement comme vital, utile ou non pertinent par rapport à un profil de l'entité. Dans le jargon de TREC, un document est dit *utile* s'il est centré sur l'entité, mais qu'il n'apporte rien par rapport à son profil existant. Cela correspond à la notion de document redondant que nous avons définie en section 1. L'annotation des documents de cette période a été faite à priori par les juges de la tâche CCR.

– une période d'évaluation allant de mars 2012 à février 2013, durant laquelle les documents doivent être analysés par les systèmes participants pour identifier leur

3. http ://trec-kba.org/trec-kba-2013.shtml
4. http ://s3.amazonaws.com/aws-publicdatasets/trec/kba/kba-streamcorpus-2013-v0_2_0-english-and-unknown-language/index.html

classe correspondante (vital, utile, non pertinent) par rapport aux requêtes entités données.

Dans ce travail, nous avons considéré toutes les requêtes pour lesquelles il existe au moins un document vital dans la période d'apprentissage et/ou d'évaluation. Ces requêtes sont au nombre de 122 : 82 personnes, 18 organisations et 22 établissements. Chaque requête est donnée sous forme d'une URL qui correspond à la page Wikipedia ou à la page Twitter de l'entité visée. Parmi l'ensemble des requêtes, 92 possèdent des documents annotés comme vitaux dans la période d'apprentissage. La mesure d'évaluation officielle adoptée dans cette tâche est la mesure $F1$. Plus précisément, la mesure F1 est la moyenne harmonique maximale entre la macro-moyenne de la précision et la macro-moyenne du rappel. Formellement $F1$ est :

$$F1 = max_i(\ F(macro\_moy\_precision@i, macro\_moy\_rappel@i)\ ) \quad [5]$$

où

F est la moyenne harmonique

$Macro\_moy\_precision@i$ est la somme des précisions de toutes les requêtes au cutoff i divisé par le nombre de requêtes

$Macro\_moy\_rappel@i$ est la somme des rappels de toutes les requêtes au cutoff i divisé par le nombre de requêtes.

4.2 Notre méthode pour l'identification de documents vitaux

Notre méthode est composée de deux étapes. D'abord, nous renvoyons les documents contenant au moins une mention de l'entité. Ensuite, nous trions ces documents selon un score de vitalité mesuré comme décrit dans la section 3.3. Pour les requêtes qui n'ont pas de documents échantillons vitaux, nous gardons le tri initial des documents obtenu à la fin de l'étape 1.

4.2.1 Étape 1 : Recherche de documents centrés sur une entité

Une entité donnée peut être mentionnée dans un document avec différentes écritures possibles appelées variantes. Afin d'identifier tous les documents mentionnant l'entité, nous devons connaître ses différentes variantes. Pour cela, pour une requête Wikipedia, nous avons exploité la page Wikipedia en extrayant le titre et les mots en gras dans le premier paragraphe comme variantes (Cucerzan, 2007). Pour une requête Twitter, nous avons exploité sa page Twitter en considérant son identifiant et son titre comme variantes. Pour chaque requête entité de l'évaluation, nous avons ensuite construit une requête composée de l'ensemble de ses variantes, chacune étant considérée comme une expression. Puis nous avons utilisé le modèle de BM25 (Robertson *et al.,* 1996) pour renvoyer les documents pertinents centrés sur l'entité.

4.2.2 Étape 2 : Trier les documents par score de vitalité

Dans cette étape nous réordonnons les documents renvoyés dans la première étape selon la mesure de vitalité décrite dans la section 3.3 (équation 4). Cette mesure traduit une certaine similarité entre un document (ou son modèle de langue) et un modèle de vitalité d'une entité θ_{Ve_x} (unidimensionnel θ_{Ve_u} ou multidimensionnel θ_{Ve_m}).

Tout le contenu d'un document ne contenant pas nécessairement des informations "vitales", nous avons considéré différentes unités d'information à évaluer :

– tout le contenu du document (**DT**).

– la concaténation des paragraphes du document mentionnant une variante de l'entité (**DP**)

– la concaténation des phrases du document mentionnant une variante de l'entité (**DS**)

Parallèlement, pour estimer le modèle de vitalité, nous avons exploité l'ensemble des documents jugés vitaux durant la période d'apprentissage. Pour chaque document vital, nous avons considéré différentes unités d'information :

– tout le contenu du document vital (**VT**).

– la concaténation des paragraphes du document vital mentionnant une variante de l'entité (**VP**)

– la concaténation des phrases du document vital mentionnant une variante de l'entité (**VS**)

Enfin, nous avons modélisé la vitalité de deux façons différentes :

– de façon unidimensionnelle (**U**) (section 3.1).

– de façon multidimensionnelle (**M**) (section 3.2).

Dans nos expérimentations, μ est fixé à 100 et nous avons pris le 30 top termes de nos modèles de vitalité (équations 1 et 2) dans l'équation 4.

4.3 Comparaison des résultats des différentes configurations de notre approche

Le tableau 1 présente les résultats des différentes configurations de notre approche générés par l'outil d'évaluation de la tâche CCR 2013[5]. La ligne $BM25$ du tableau correspond à notre baseline, c'est à dire les résultats à l'issue de la première étape de notre approche. Pour chaque configuration du tableau, nous indiquons entre parenthèses le cutoff correspondant au résultat.

Les résultats du tableau 1 montre que le fait de ne considérer que les phrases des documents échantillons vitaux (D ?-VS-*) pour l'estimation du modèle vital, permet d'améliorer peu ou pas le tri des documents vitaux par rapport à la baseline (BM25).

5. https ://github.com/trec-kba/kba-scorer

Runs	unidimensionnelle (U)	multidimensionnelle (M)
DS-VS	0.352 (100)	0.345 (90)
DS-VP	0.354 (90)	0.348 (90)
DS-VT	0.355 (130)	0.352 (110)
DP-VS	0.347 (70)	0.342 (120)
DP-VP	0.356 (30)	0.350 (30)
DP-VT	0.350 (110)	0.348 (110)
DT-VS	0.346 (80)	0.344 (80)
DT-VP	0.353 (30)	0.347 (80)
DT-VT	**0.383** (30)	0.369 (30)
BM25	0.345	

Tableau 1. *Résultats des différentes configurations de notre système selon la mesure officielle F1 (Eq. 5)*

En prenant des unités plus grandes, les paragraphes (D ?-VP-*), les résultats sont meilleurs que les phrases. Les meilleurs résultats sont obtenus en prenant le document entier. Nous illustrons dans la figure 1 l'impact de l'unité "vitale" choisie sur les résultats.

Nous voyons aussi que le choix des unités d'information à considérer dans les documents de test est important. On constate que prendre en compte des granularités similaires pour les documents échantillons et les documents vitaux (DT-VT-*), permet de mieux cerner la vitalité comme le montre la figure 2.

Enfin, la modélisation unidimensionnelle de la vitalité (DT-VT-U) en considérant la concaténation de tous les documents échantillons vitaux de l'entité, semble être meilleure qu'une modélisation multidimensionnelle (DT-VT-M) qui estime un sous-modèle de vitalité pour chaque document vital. Ceci est valide pour tous les points de cutoff comme le montre la figure 3.

Nous constatons que ces résultats sont très proches. Ceci peut être expliqué par deux raisons. D'une part, ces configurations gardent le tri initial des documents (Étape 1) pour les 30 requêtes pour lesquelles nous ne disposons pas des documents vitaux d'entraînement. Une autre explication peut venir de la mesure F1 qui conserve le point de *cutoff* pour lequel la F-mesure des macro-moyennes est la plus élevée (Eq. 5).

Nous remarquons dans les figures 1, 2 et 3, que toutes les configurations atteignent le même score au cutoff 1000 puisqu'elles retrient les même documents obtenus à l'issue de l'étape 1. La différence ne peut être vue qu'au niveau des top premiers documents restitués. Afin de visualiser de façon plus claire les performances de ces configurations, nous illustrons dans le tableau 2 les résultats obtenus pour un cutoff fixé à 40 qui représente le nombre moyen de documents vitaux par requête.

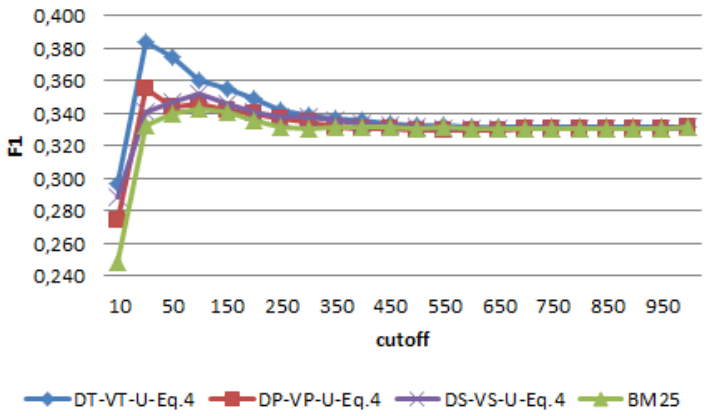

Figure 1. *Impact de l'unité d'information considérée (S/P/T) dans l'estimation du modèle vital*

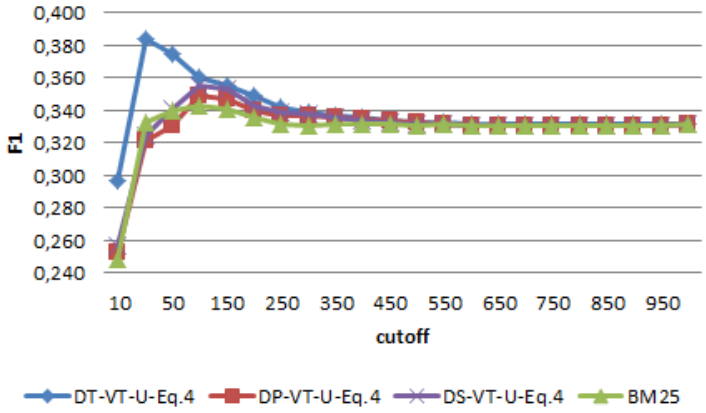

Figure 2. *Impact de l'unité d'information considérée dans le document (S/P/T) dans l'évaluation de sa vitalité*

Dans la section suivante, nous positionnons notre approche par rapport aux méthodes de l'état de l'art.

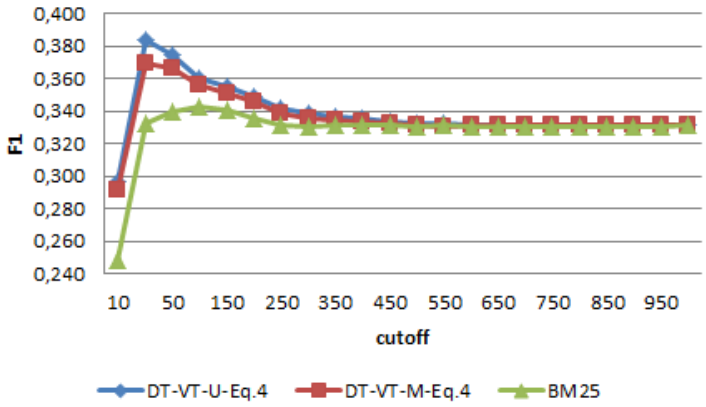

Figure 3. *Variation des performances de notre approche en considérant tout le contenu des documents (DT-VT), en fonction de la modélisation (unidimensionnelle U ou multidimensionnelle M)*

Runs	unidimensionnelle (U)	multidimensionnelle (M)
DS-VS	0.343	0.339
DS-VP	0.352	0.343
DS-VT	0.334	0.327
DP-VS	0.338	0.334
DP-VP	0.351	0.343
DP-VT	0.328	0.324
DT-VS	0.342	0.343
DT-VP	0.352	0.343
DT-VT	0.379*	0.368*
BM25	0.335	

Tableau 2. *Résultats de notre système en fixant le cutoff à 40 (F-mesure de la macro moyenne du rappel et la macro moyenne de la précision calculées pour un cutoff fixé à 40). Une astérisque indique que l'amélioration est statistiquement significative par rapport à la baseline selon le test t de Student pairé et bilatéral avec p <0.05*

4.4 Comparaison de notre approche par rapport aux méthodes de l'état de l'art

Le tableau 3 compare notre meilleure configuration par rapport aux trois meilleurs systèmes de la tâche CCR 2013. Le meilleur système (F1=0.360) représenté dans le tableau par **S1** se base sur une classification supervisée utilisant cinq familles de cri-

Méthode	F1
BM25 (baseline)	0.345 $(R = 0.638, P = 0.237)$
DT-VT-U	**0.383** $(R = 0.500, P = 0.311)$
S1 (état de l'art)	0.360 $(R = 0.601, P = 0.257)$
S2 (état de l'art)	0.316 $(R = 0.591, P = 0.216)$
S3 (état de l'art)	0.309 $(R = 0.695, P = 0.199)$
Médiane CCR 2013	0.201
Moyenne CCR 2013	0.193
Min CCR 2013	0

Tableau 3. *Comparaison de notre notre système par rapport aux systèmes participants de la tâche CCR 2013. La mesure officielle de comparaison est la $F1$ (Eq. 5)*

tères (Wang *et al.,* 2013). Le second meilleur système **S2** utilise l'expansion de la requête par des mentions d'entités reliées. Le troisième meilleur système **S3** se base essentiellement sur l'appariement simple des mentions de l'entité.

Notre meilleure configuration (DT-VT-U-Eq.4) dépasse les systèmes participants de la tâche, avec une mesure de $F1 = 0.383$.

Nous remarquons qu'en adoptant la mesure officielle de la tâche F1 (Eq. 5), renvoyer tous les documents qui mentionnent au moins une variante de l'entité comme dans la méthode S3 ou aussi dans notre baseline, permet d'avoir de bons résultats dans la tâche grâce à un bon taux de rappel. Cependant la précision de ces méthodes reste faible.

Comme les résultats détaillés (en fonction des points de cutoff) de tous les systèmes participants ne sont pas rendus publics, il est difficile de faire des comparaisons plus approfondies.

5 Conclusion et perspectives

Dans cet article, nous nous sommes intéressés à la recherche de documents vitaux permettant la mise à jour d'un profil d'une entité donnée. Nous nous sommes basés sur les modèles de langue pour estimer un modèle permettant la génération de documents vitaux pour une entité donnée. Nous avons évalué notre approche dans le cadre de la tâche CCR de TREC KBA 2013. Nous avons montré que la vitalité est mieux estimée en considérant tout le contenu du document.

Ce travail ouvre beaucoup de perspectives intéressantes. A court terme, nous souhaitons estimer nos modèles de vitalité de façon globale, indépendamment de la présence d'un ensemble d'apprentissage pour l'entité. Nous pensons également à combiner notre modèle de vitalité avec d'autres critères qui peuvent servir à détecter la vitalité d'un document comme le "buzz" des documents parlant d'une information nou-

velle, les contenus générés par l'utilisateur dans les réseaux sociaux, etc. La première partie de notre approche, permettant de renvoyer les documents centrés sur l'entité, peut également être améliorée.

A terme, nous pensons enfin exploiter les documents vitaux pour extraire automatiquement des informations vitales sur une entité, permettant ainsi la mise à jour de son profil. Par exemple l'information "Dernier album : Loved Me Back To Life" peut servir à enrichir ou mettre à jour l'*infobox* de l'article Wikipedia de l'entité "Céline Dion".

6 Bibliographie

Abbes R., Pinel-Sauvagnat K., Hernandez N., Boughanem M., « IRIT at TREC Knowledge Base Acceleration 2013 : Cumulative Citation Recommendation Task », *Notebook of the TExt Retrieval Conference 2013 (TREC 2013)*, Gaithersburgh, MD, USA., 2013.

Aggarwal N., Buitelaar P., « Query Expansion Using Wikipedia and Dbpedia. », *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

Bonnefoy L., Bouvier V., Bellot P., « A weakly-supervised detection of entity central documents in a stream », *SIGIR*, p. 769-772, 2013.

Bron M., Balog K., de Rijke M., « Ranking Related Entities : Components and Analyses », *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, ACM, New York, NY, USA, p. 1079-1088, 2010.

Cucerzan S., « Large-Scale Named Entity Disambiguation Based on Wikipedia Data », *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Association for Computational Linguistics, Prague, Czech Republic, p. 708-716, June, 2007.

Dietz L., Dalton J., « Umass at TREC KBA 2013 », *Notebook of the TExt Retrieval Conference 2013 (TREC 2013)*, Gaithersburgh, MD, USA., 2013.

Frank J. R., Bauer S. J., Kleiman-Weiner M., Roberts D. A., Tripuraneni N., Zhang C., Re C., « Evaluating Stream Filtering for Entity Profile Updates for TREC 2013 », *Notebook of the TExt Retrieval Conference 2013 (TREC 2013)*, Gaithersburgh, MD, USA., 2013.

Frank J. R., Kleiman-Weiner M., Roberts D. A., Niu F., Zhang C., Re C., Soboroff I., « Building an Entity-Centric Stream Filtering Test Collection for TREC 2012 », *Proceedings of the Text REtrieval Conference (TREC)*, 2012.

Harman D., « Overview of the TREC 2002 Novelty Track », *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002), NIST Special Publication 500-251*, p. 46-55, 2002.

Larkey L. S., Allan J., Connell M. E., Bolivar A., Wade C., « UMass at TREC 2002 : Cross language and novelty tracks », *Notebook Proceedings of TREC 2003*, p. 721-732, 2002.

Lavrenko V., Croft W. B., « Relevance Based Language Models », *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, ACM, New York, NY, USA, p. 120-127, 2001.

Li X., Li C., Yu C., « Entity-Relationship Queries over Wikipedia », *ACM Trans. Intell. Syst. Technol.*, vol. 3, n° 4, p. 70 :1-70 :20, September, 2012.

Liu X., Fang H., « A Related Entity based Approach for Knowledge Base Acceleration », *Notebook of the TExt Retrieval Conference 2013 (TREC 2013)*, 2013.

Ponte J. M., Croft W. B., « A Language Modeling Approach to Information Retrieval », *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, ACM, New York, NY, USA, p. 275-281, 1998.

Pound J., Hudek A. K., Ilyas I. F., Weddell G., « Interpreting Keyword Queries over Web Knowledge Bases », *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, ACM, New York, NY, USA, p. 305-314, 2012.

Pound J., Mika P., Zaragoza H., « Ad-hoc object retrieval in the web of data », *Proceedings of the 19th international conference on World wide web*, WWW '10, ACM, New York, NY, USA, p. 771-780, 2010.

Robertson S., Walker S., Jones S., Hancock-Beaulieu M., Gatford M., « Okapi at TREC-3 », *Text REtrieval Conference (TREC)*, p. 109-126, 1996.

Suzuki Y., Yoshikawa M., « Assessing quality score of Wikipedia article using mutual evaluation of editors and texts », *Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '13, p. 1727-1732, 2013.

Wang J., Song D., Lin C.-Y., Liao L., « BIT and MSRA at TREC KBA CCR Track 2013 », *Notebook of the TExt Retrieval Conference 2013 (TREC 2013)*, Gaithersburgh, MD, USA., 2013.

Zhang C., Xu W., Liu R., Zhang W., Zhang D., Ji J., Yang J., « PRIS at TREC KBA 2013 », *Notebook of the TExt Retrieval Conference 2013 (TREC 2013)*, Gaithersburgh, MD, USA., 2013.

Zhang Y., Callan J., Minka T., « Novelty and Redundancy Detection in Adaptive Filtering », *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, ACM, New York, NY, USA, p. 81-88, 2002.