Vers une détection en temps réel de documents Web centrés sur une entité donnée

Ludovic Bonnefoy[1,2], **Vincent Bouvier**[3,4], **Romain Deveaud**[1], **Patrice Bellot**[3]

[1] *Université d'Avignon - CERI/LIA* [2] *iSmart*
ludovic.bonnefoy, romain.deveaud@alumni.univ-avignon.fr
[3] *Université d'Aix-Marseille - LSIS* [4] *Kware*
{vincent.bouvier, patrice.bellot}@lsis.org

RÉSUMÉ. *La tâche de désambiguïsation des entités nommées consiste à lier une mention ambiguë d'une entité dans un document à l'entité correspondante dans une base de connaissances. Dans ce travail, nous nous plaçons dans un cadre applicatif "inverse" et nous ajoutons une contrainte temporelle : nous souhaitons surveiller un flux de nouveaux documents Web et déterminer quels sont ceux mentionnant une entité donnée tout en mesurant l'importance de l'information contenue. Une telle approche peut servir à recommander des documents à des contributeurs si une information mérite d'être ajoutée dans la base de connaissances cible. Notre approche repose sur l'utilisation de deux classifieurs prenant en compte, pour déterminer l'intérêt d'un document du flux, des indices comme la fréquence de mentions de l'entité dans le temps ou dans le document, sa position ou encore la présence d'entités liées connues. Notre approche et l'impact des paramètres utilisés ont été évalués via une participation à la tâche "Knowledge Base Acceleration" de TREC 2012 et a positionné notre équipe au rang 3 sur 11 (Bonnefoy* et al., *2012).*

ABSTRACT. *Name entity disambiguation is the task of linking an ambiguous name in a document to the unique real-world entity in a kwnoledge base (KB) its represents. We took the opposite problem and add a time constraint : we monitor a data stream to detect in real-time documents about an entity from a KB and determine to what extent the information in those documents matter. It could be used to reduce time lag between the moment a new important information about an entity shows up and the moment it is added to the knowledge base. We used Random Forests combined with time-related features (eg. count of mentions in time) and document and related entities centric features to tackle this problem. The effectiveness and impact of the features used have been evaluated through our participation to the "Knowledge Base Acceleration" task at TREC 2012 and positionned our team rank 3 on 11 (Bonnefoy* et al., *2012).*

MOTS-CLÉS : *entité nommée, base de connaissances, kba, trec, flux*

KEYWORDS: *named entities, knowledge base, kba, trec, stream*

1. Introduction

Les entités nommées sont au cœur de nombreux travaux dans le domaine de la recherche d'information et du traitement de la langue naturelle écrite. Cet intérêt a été impulsé et maintenu par de multiples campagnes d'évaluation leur ayant accordé une part importante : MUC (*Named Entity task*[1]), TREC (avec la tâche *Question Answering* (Voorhees, 1999)), etc.

Les premières méthodes non supervisées de recherche d'entités nommées étaient basées sur des ensembles de patrons d'extraction (Nadeau *et al.,* 2007) et, aujourd'hui encore, il est conseillé de procéder de la sorte si un corpus d'entraînement n'est pas disponible pour les types souhaités (Sekine *et al.,* 2004). Avec l'arrivée des premiers corpus d'apprentissage pour quelques types (personne, lieu, organisation et date) des approches supervisées sont apparues avec l'utilisation des modèles de Markov cachés (Bikel *et al.,* 1997), des arbres de décision (Sekine, 1998) ou encore des SVMs (Asahara *et al.,* 2003) et des CRFs (McCallum, 2003). Des méthodes faiblement (ou semi-)supervisées ont aussi été étudiées telle que le *bootstrapping* en exploitant différents critères comme les relations syntaxiques (Cucchiarelli *et al.,* 2001) ou synonymiques (Pasca *et al.,* 2006).

Aujourd'hui, le domaine a atteint une certaine maturité mais les performances stagnent quelque peu malgré les importants progrès restant à faire (comme par exemple la capacité à gérer tous type d'entités aussi fins que souhaité et pas seulement les types de très haut niveau). Les travaux sur le sujet, bien que nombreux, se concentrent désormais sur des sous problèmes présentant des caractéristiques et difficultés très spécifiques telles que la reconnaissance des entités dans le domaine biomédical (très difficile à segmenter par leur formes) (Atkinson *et al.,* 2012) ou dans les tweets (très peu de contexte) (Liu *et al.,* 2012).

Cependant, de nouvelles thématiques de recherche ont émergé. Parmi elles, deux nous intéressent particulièrement : la résolution des co-références des entités nommées au sein d'un document ou entre plusieurs documents ainsi que la tâche *Entity Linking* (EL) de TAC (*Text Analysis Conference*). Ces deux tâches répondent au problème soulevé par l'ambiguïté des noms des entités : une même entité peut être désignée par plusieurs mots différents et, à l'inverse, un même mot peut désigner plusieurs entités.

La première tâche correspond à lier entre elles, dans un document ou dans une collection, différentes expressions ayant la même référence dans un contexte unique : par exemple "Elvis" et "The King" (ce dernier terme peut référer à Elvis Presley mais aussi, bien sûr, à d'autres personnes dans un contexte non musical). Les approches existantes se distinguent d'une part par leur niveau de supervision mais aussi par le niveau auquel sont estimés les différents paramètres pour la prise de décision : de nombreuses approches résolvent les co-références en analysant les expressions par paires tandis que d'autres choisissent un angle plus large et travaillent sur l'ensemble des mentions à la fois. Dans tous les cas, les critères utilisés sont des éléments classiques :

1. http ://cs.nyu.edu/faculty/grishman/NEtask20.book_1.html

critères lexicaux (par exemple le recouvrement lexical entre les deux occurrences), sémantiques (mise en relation via des ressources terminologiques telles que Wordnet) ou encore syntaxiques (structures syntaxiques dans lesquelles apparaissent les multiples occurrences). Un état de l'art du domaine est proposé par (Clark *et al.,* 2008).

La seconde tâche, *Entity Linking*, se propose de désambiguïser les mentions d'entités rencontrées dans un document en les liant à des entités mentionnées dans une base de connaissances. Cette tâche a vu le jour grâce à l'émergence de bases de connaissances et d'encyclopédies de plus en plus complètes telles que Wikipédia[2] ou Freebase[3]. Outre la désambiguïsation en elle-même, cette tâche présente d'autres intérêts tant du côté des interfaces utilisateurs (par l'ajout de liens explicites des occurrences vers les entités elles-mêmes) que du côté des systèmes automatiques comme première étape vers une population automatique des bases de connaissances.

Cet axe de recherche est soutenu par la campagne d'évaluation *Text Analysis Conference* (TAC)[4] avec la tâche *Entity Linking* au sein de la piste *Knowledge Base Population* (KBP) qui met à disposition plusieurs milliers de *topics* (couple "mention d'entité et document de presse associé") ainsi que les jugements de pertinence associés, une base de connaissances (créée à partir de Wikipédia) et un corpus fort de plusieurs centaines de milliers de documents journalistiques, de pages Web et de blogs. Les approches implémentées par les participants s'apparentent en général à celle des méthodes de résolution des co-références, combinées ou non à l'utilisation de techniques de retour de pertinence et d'expansion de requêtes (Artiles *et al.,* 2011). De plus, la résolution de cette tâche nécessite de déterminer si l'occurrence d'une entité correspond ou non à une entité déjà présente dans la base de connaissances. Pour cela, des mesures de similarité entre les entrées de la base et le contexte de l'occurrence ainsi que des mesures surfaciques entre les mots sont utilisées. Le compte rendu de la tâche KBP à TAC 2011 (Ji *et al.,* 2011) présente un nombre important d'approches ainsi qu'une évaluation approfondie des performances obtenues. Les approches les plus performantes sont majoritairement supervisées et permettent d'estimer automatiquement le poids de nombreux critères qu'ils soient syntaxiques ou lexicaux.

Récemment, des chercheurs proposèrent de considérer la tâche à l'envers, en partant d'entités dans une base de connaissances et cherchant, dans une collection, les documents les mentionnant. Cette tâche ne serait qu'une nouvelle tâche ad hoc supplémentaire si une nouvelle contrainte forte n'avait été ajoutée : le temps. La tâche consiste donc à surveiller un flux de documents et à déterminer lesquels concernent les entités cibles. Les premiers travaux se sont orientés vers l'exploitation des réseaux sociaux, en particulier Twitter, avec pour objectif de pouvoir suivre l'évolution en direct d'évènements nommés, comme les catastrophes naturelles (Lee, 2012). Cette contrainte de réactivité et de prise de décision en temps réel apporte de nouveaux défis. En effet, de nombreuses approches utilisées pour la désambiguïsation des mentions d'entités ont recours à des ressources externes ou à des traitements coûteux (telle

2. http ://www.wikipedia.org/
3. http ://www.freebase.com/
4. http ://www.nist.gov/tac/2012/KBP/index.html

qu'une analyse grammaticale ou une résolution des anaphores) et ne sont pas applicables ici. De plus, au cours du temps les entités changent (aussi bien leurs caractéristiques que les valeurs de ces dernières, mais aussi la façon de les dénommer) et leur représentation initiale, si elle n'est pas mise à jour, peut devenir obsolète et amener à ignorer de nouvelles données

L'utilisation de sources telles que Twitter est problématique en particulier à cause de la taille des messages qui oblige à ne travailler qu'avec des entités très populaires, présentes dans de gros volumes de tweets (Davis *et al.,* 2012). En 2012, une nouvelle piste TREC, nommée *Knowledge Base Acceleration* (KBA)[5] a été créée. Cette piste pose le problème suivant : dans une importante base de connaissances comme Wikipédia, il y a plus d'entités qu'il n'y a de contributeurs pour la mettre à jour. Cela engendre un temps de latence médian de 356 jours (Frank *et al.,* 2012) avant qu'un élément nouveau, en rapport avec une entité donnée, ne soit reporté sur sa page. Ne pourrait-on pas raccourcir ce délai en soumettant à des contributeurs en charge d'une entrée de la base tous les nouveaux documents mentionnant l'entité et ordonnés en fonction d'une estimation de leur importance. La tâche KBA de TREC 2012 consiste ainsi à trouver, pour 29 entités de Wikipédia choisies pour leur ambiguïté, tout document les mentionnant au sein d'une large collection et à attribuer à ces derniers une classe d'importance : non pertinent (bien que le document mentionne l'entité, aucune information précise la concernant n'est présente), pertinent ou central. La notion de document central est capitale dans cette tâche car c'est la capacité des systèmes à trouver les documents de ce type qui est mesurée. Un document est considéré comme central si il apporte une information majeure sur l'entité et devrait absolument figurer dans l'entrée correspondante (ici sa page Wikipédia). Par exemple, pour l'entité *Barack Obama*, un document mentionnant une visite diplomatique est pertinent mais non central tandis qu'un document discutant de sa réélection l'est. Notons que la question de la nouveauté de l'information a été mise de côté pour l'édition 2012 de KBA.

Cette tâche a de fortes similarités avec la tâche *Filtering* organisée à TREC à la fin des années 90 (Robertson *et al.,* 2002a). Celle-ci était définie comme suit : à partir d'une requête utilisateur ainsi que d'un ensemble de documents jugés pertinents, déterminer, pour chaque nouveau document apparaissant dans un flux, s'il répond ou non aux attentes de l'utilisateur. Nombre d'approches eurent recours à des techniques traditionnelles en recherche d'information (par ex. Okapi (Robertson *et al.,* 2002b), Rocchio (Collins-Thompson *et al.,* 2002)...) pour associer un score aux documents. Un seuil était ensuite calculé à partir du jeu de documents donné et d'un corpus d'apprentissage. Pour la majorité des autres approches performantes, des SVMs furent utilisés (par exemple (Cancedda *et al.,* 2002)) principalement selon des critères lexicaux (comme le compte des n-grammes (Mihalcea, 2002)). Cependant, des différences importantes existent et rendent difficiles la réutilisation de ces approches pour KBA : la taille de la collection était cent fois plus petite, elle ne contenait que des documents journalistiques (le corpus de KBA contient aussi des documents provenant de blogs et des pages Web), l'unité de temps considérée était la journée (contre l'heure) –elle per-

5. http ://trec-kba.org/

mettait d'avoir plus de recul sur un évènement– et, enfin, à chaque prise de décision les systèmes étaient informés de la classe associée au document par les annotateurs (cela permettait de ré-estimer dynamiquement les modèles et de réduire la divergence du modèle initial au fil du temps).

Malgré (ou à cause de) toutes ces difficultés supplémentaires, les meilleurs systèmes de KBA 2012 sont des approches relativement simples. (Kjersten *et al.,* 2012) utilisèrent un classifieur SVM basé sur les mots et les entités nommés rencontrées comme composante vectorielle tandis que (Araujo *et al.,* 2012) considérèrent un document comme central à partir du moment où il contenait l'entité telle quelle.

Notre approche, troisième parmi celles de onze équipes, est basée sur l'utilisation de classifieurs pour déterminer le degré d'intérêt d'un document. Notre but était de proposer et d'étudier quels indices caractérisent un document "central". Les éléments étudiés appartiennent à trois catégories : prise en compte du temps, caractéristiques des occurrences des mots-entités dans le document et présence d'entités liées.

Cet article se décompose de la manière suivante : notre approche est présentée dans une première partie (choix du classifieur et des critères utilisés), puis nous analysons et commentons les résultats obtenus dans le cadre de TREC KBA 2012.

2. Détection de documents centrés sur une entité dans un flux de documents

Considérons un flux de documents issus du Web et $v_1, v_2, ..., v_N \in V_e$ les différentes écritures (variantes) d'une entité donnée e. Notre tâche est d'être en mesure de déterminer si un nouveau document du flux contenant une occurrence v_i fait bien référence à e et s'il contient des informations importantes à son sujet. Notre approche repose sur l'utilisation de deux classifieurs en cascade permettant de déterminer si un document est pertinent ou non mais aussi pour distinguer deux niveaux de pertinence : intéressant et central (qui mériterait d'être cité dans la base de connaissances).

2.1. *Pré-traitements*

Dans notre travail, à l'instar de la tâche KBA, la "base de connaissances" considérée est Wikipédia. Une entrée Wikipédia ne contient pas explicitement (dans les *infoboxes*) les différentes écritures v_i liées à l'entité. Cependant, obtenir ces variantes peut se révéler important pour améliorer le rappel d'un système (par exemple ne pas exclure de l'analyse un document mentionnant *The King* pour une recherche portant sur *Elvis Presley*). Nous avons procédé à la manière de (Cucerzan, 2007) en considérant comme variantes les éléments en gras dans le premier paragraphe de l'entrée Wikipédia, le nom des pages de redirection et le texte des liens pointant vers celles-ci.

Boris Berezovsky (homme d'affaire)
boris berezovsky
boris abramovich berezovsky

Boris Berezovsky (pianiste)
boris berezovsky
boris vadimovich berezovsky

Figure 1. *Variantes trouvées pour Boris Berezovsky l'homme d'affaire et le pianiste*

Un poids est ensuite associé à chaque variante :

$$w(v_i) = \begin{cases} 1 \text{ si } v_i \text{ correspond au nom de l'entrée} \\ \frac{tf(v_i)}{\sum_{v_j \in V_e} tf(v_j)} \text{ sinon} \end{cases} \quad [1]$$

2.2. *Classification*

Nous avons utilisé des forêts d'arbres de décision souvent aussi performants que des SVMs ou autres classifieurs de l'état de l'art mais qui, à l'instar des arbres de décisions isolés (bien que dans une moindre mesure) permettent une analyse explicite de l'importance relative des critères utilisés pour la classification (Breiman, 2001).

Comme dans tout problème de classification, le choix des critères est central. De nombreux travaux décrivent un nombre important de critères susceptibles de caractériser un document Web (Qi *et al.,* 2009) mais, bien que la plupart soient utilisables ici, nous avons décidé de nous focaliser uniquement sur les éléments spécifiques au problème en évitant la multiplication de critères parfois complexes et coûteux à estimer. Les critères étudiés dans ce travail appartiennent à trois catégories : prise en compte du temps, de caractéristiques des occurrences des variantes dans le document et des relations de l'entité avec d'autres entités.

2.2.1. *Analyse temporelle*

La prise en compte du temps doit nous renseigner sur l'importance de l'entité dans la période de parution d'un document et nous permettre d'estimer si la période en question est propice à l'apparition de documents intéressants. Plusieurs mesures sont proposées pour quantifier cette importance.

Nombre de documents sur 24 heures : Le premier critère considéré est le nombre de documents contenant une mention de l'entité dans les précédentes 24 heures. Nous souhaitons ainsi mesurer l'apparition d'une nouvelle tendance.

Nombre de documents sur 7 jours : Cependant il est difficile d'estimer une évolution sans prendre en compte une échelle de temps plus large. Cette même mesure

est donc aussi effectuée sur 7 jours.

Moyenne et écart type du nombre de documents par jour : Afin de refléter le caractère commun ou exceptionnel du nombre de documents nous utilisons comme critère le nombre moyen de documents par jour les 7 derniers jours ainsi que l'écart type.

Nombre de titres avec une mention sur 7 jours : Le titre est intuitivement un indicateur fort de l'orientation d'un document et la mention de l'entité dans un grand nombre de titres les 7 derniers jours semble être un indicateur fort.

Probabilité d'avoir un document pertinent sachant les n jours précédents : Nous souhaitions par delà déterminer la probabilité de trouver un document pertinent sachant les observations faites pour les jours précédents. Pour cela des pseudo n-grammes de documents sont utilisés dont chaque caractère renseigne sur l'observation faite pour un jour donné. Il peut prendre la valeur 1 si au moins un document pertinent a été trouvé ce jour là, et 0 sinon (le n-gramme 001 signifie que le jour précédent, au moins un document pertinent a été trouvé, et les deux jours précédents, aucun). Dans nos expériences, n peut prendre la valeur 1 ou 2.

2.2.2. *Analyse du contenu des documents*

Les critères précédents nous indiquent si l'entité a une actualité forte au moment de l'apparition du document mais ne caractérisent pas le document en lui-même (tous les documents du même jour obtiennent les mêmes valeurs pour ces critères).

Nombre de mentions dans le document : Un document centré sur une entité doit la mentionner de nombreuses fois. Le premier critère est égal à la somme des poids de chaque variante (cf section 2.1) de l'entité rencontrée dans le document :

$$s(d) = \sum_{v_i \in V_e} tf(v_i, d) \times w(v_i) \quad [2]$$

où $tf(v_i, d)$ est le nombre d'occurrences de v_i dans d normalisé par rapport à la taille de d.

Position des mentions : Tout comme pour la création automatique de résumés (Das *et al.,* 2007), la position des occurrences dans le texte semble être importante (notamment dans le titre, le début et la fin d'un document). Dans cet esprit nous avons un ensemble de critères correspondants au nombre de mentions par tranche de 10% du texte (en nombre de mots) et un autre pour le titre.

$$s(d_t) = \sum_{v_i \in V_e} tf(v_i, d_t) \times w(v_i) \quad [3]$$

où d_t correspond à une sous-partie du document (le titre ou une tranche de 10%) et $tf(v_i, d)$ au nombre d'occurrences de v_i dans d normalisé par rapport à la taille de d.

Similarité vectorielle : Les deux derniers critères correspondent au calcul d'une similarité entre le document et l'entrée de l'entité dans la base (ici sa page Wikipédia). La similarité utilisée est la similarité cosinus et les vecteurs sont composés des unigrammes présents dans les documents pour le premier, et des bigrammes pour le second.

2.2.3. *Étude des entités liées*

La présence dans un document d'une mention de l'entité et d'une ou plusieurs des entités qui lui sont liées dans la base de connaissances nous a semblé un facteur supplémentaire fortement discriminant pour déterminer à quel point un document est centré sur l'entité.

Nous avons extrait les entités liées de deux manières à partir d'une entrée de Wikipédia en considérant :

– toutes les entités trouvées dans la page à l'aide d'un outil de reconnaissance des entités nommées (dans notre cas le Stanford NER[6]) ;

– les entités pointées par la page (peut être plus importantes que les autres).

Pour chaque entité liée re_i à e un poids est associé selon :

$$w(re_i, e) = \frac{tf(re_i, e)}{\sum_{re \in RE(e)} tf(re, e)} \quad [4]$$

où RE est l'ensemble des entités liées à l'entité e.

Mesure de la présence d'entités liées : Ce critère reflète la présence de ces entités dans le document en prenant en compte l'importance de la relation :

$$SEL(d, e) = \sum_{re_i j \in RE} w(re_i, e) \quad [5]$$

2.2.4. *Détection des documents centrés sur une entité dans un flux*

Pour déterminer si un nouveau document du flux est intéressant au regard d'une entité donnée, deux classifieurs de type forêt d'arbres de décision construits selon les critères précédents sont utilisés séquentiellement. Le premier pour sélectionner les documents pertinents, et le second pour distinguer, parmi les documents pertinents, ceux qui sont très intéressants (classe "central") des autres ("intéressants").

3. Cadre expérimental et résultats

Dans cette section nous commençons par présenter le cadre d'évaluation proposé par la tâche KBA à TREC 2012 et les adaptations de notre système, puis, dans un second temps, les résultats officiels seront analysés.

3.1. *TREC KBA 2012*

Dans le cadre de cette campagne d'évaluation, un nouveau corpus a été élaboré. Il est composé de trois catégories de documents pour un total de près de 9To de don-

6. http ://nlp.stanford.edu/ner/index.shtml

nées pour environ 500 millions de documents (*cf.* tableau 1). Les trois catégories de documents sont :

– **Social :** ensemble de documents provenant de blogs et forums ;

– **Web :** documents web provenant de la base de Bitly[7] ;

– **Presse :** documents journalistiques.

Ces documents ont été collectés entre octobre 2011 et avril 2012. A chaque document sont associées la date et l'heure précise de sa publication. Le corpus est divisé en deux parties : d'octobre à décembre 2011 pour l'entraînement, et de janvier à avril 2012 pour le test (évaluation officielle).

	Presse	Web	Social
# docs	134 625 663	5 400 200	322 650 609
taille	8072GB	350GB	531GB

Tableau 1. *Corpus KBA 2012 : nombre de documents et taille par catégorie.*

Pour l'évaluation, 29 entités ont été sélectionnées pour leur difficulté et leur degré d'ambiguïté (voir la liste figure 3) ainsi que leur faible nombre d'occurrences dans le corpus (l'objectif étant de ne pas chercher à cibler des entités trop populaires ni, à l'inverse, des personnes pour lesquelles trop peu d'information sont disponibles sur le Web). Cela a permis, *a priori* , l'annotation manuelle de la quasi-totalité des documents les mentionnant (le rappel étant estimé à 91% après analyse des résultats renvoyés par les participants (Frank *et al.,* 2012)). Les annotateurs devaient associer l'une des trois classes suivantes aux documents : non pertinent, intéressant, central. La mesure officielle est la F-mesure (moyenne harmonique entre précision et rappel).

3.2. *Associer un score aux documents*

L'évaluation dans le cadre de KBA impose aux systèmes de retourner une liste des documents triés pour en mesurer les performances. Pour chaque entité, les participants devaient retourner une liste de documents triés selon leur score alors que notre approche, telle que présentée en section 2, associe une étiquette à chaque document parmi "non pertinent", "intéressant" et "central". Cependant, les classifieurs utilisés associent à chaque prise de décision un score entre 0 et 1 faisant office de score de confiance. Utilisant cette information, nous procédons de la sorte pour associer un score à un document :

$$S_1(d_i) = \begin{cases} s(d_i, c_{np,p}) \times s(d_i, c_{i,c}) \text{ si } s(d_i, c_{np,p}) >= 0,5 \\ \text{retiré de la liste sinon} \end{cases} \quad [6]$$

7. https ://bitly.com/

où le score du document d_i est donné comme le produit du score renvoyé par le classifieur $c_{np,p}$ (départageant les non pertinents (np) des pertinents (p)) et par le classifieur $c_{i,c}$ (associant "intéressant" (i) ou "central" (c)).

Les classifieurs commettant des erreurs et afin de ne pas pénaliser un document mal classé par $c_{np,p}$ (qui serait donc exclu de la liste), une deuxième fonction de score est considérée :

$$S_2(d_i) = \begin{cases} s(d_i, c_{np,p}) \text{ si } s(d_i, c_{np,p}) < 0.5 \\ 0.5 + \frac{s(d_i, c_{np,p}) \times s(d_i, c_{i,c})}{2} \text{ sinon} \end{cases} \quad [7]$$

Deux soumissions (*runs*), correspondant à ces deux fonctions de score, ont été proposées pour notre participation à la tâche KBA 2012.

3.3. *Résultats*

Dans cette sous-section, sont présentés les résultats officiels pour notre participation à la tâche KBA à TREC 2012. L'évaluation porte sur la capacité des systèmes à trouver dans le flux les documents dits "centraux".

La figure 2 présente les résultats moyens de toutes les soumissions des participants. Ces résultats globaux nous permettent d'observer que notre approche se place en 3e position. De plus, les scores obtenus (0,342 pour notre run 1 et 0,33 pour notre run 2) situent nos propositions très près du meilleur système (0,359) et nettement au-dessus des scores médian (0,289) et moyen (0,22). Malgré cela, les informations reportées dans le tableau 2 nous obligent à relativiser ce succès. Elles représentent les proportions de documents, mentionnant ou non une entité, pertinents ou non dans les premières semaines de la plage temporelle des documents du corpus. Un système capable de récupérer l'ensemble des documents mentionnant une entité obtiendrait une F-mesure de 0,377 et serait en première position... (ces résultats triviaux sont probablement généralisables sur l'ensemble du corpus même si cela nécessite confirmation). Ce résultat souligne la complexité de la tâche dont l'une des principales difficultés est d'être en mesure de trouver toutes les variantes d'une entité afin d'obtenir un rappel élevé.

	non pertinent	intéressant	central
contient une mention	11853	13971	7806
zéro mention	15530	61	0

Tableau 2. *Statistiques estimées sur les premières semaines du corpus exprimant les proportions des catégories associées à chaque document contenant ou non une mention d'une entité.*

La figure 3 présente une analyse de nos résultats entité par entité. Nos deux soumissions y sont présentées (barres verticales) ainsi que la médiane et un score re-

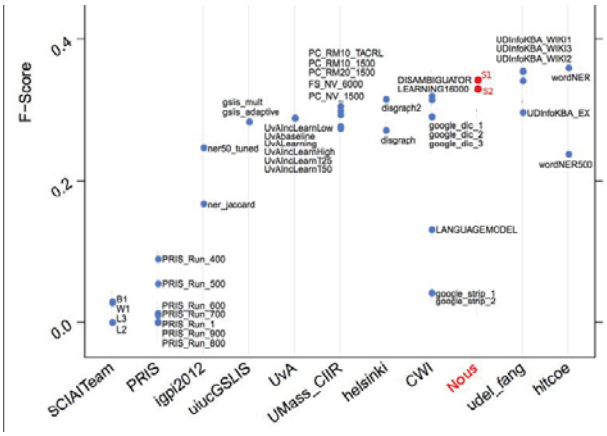

Figure 2. *Résultats officiels pour les documents "centraux" uniquement à KBA 2012 pour la F-mesure.*

présentant la facilité de l'entité. Cette mesure est fonction du nombre de documents annotés comme pertinents et non pertinents (pour rappel ont été annotés tous les documents contenant une mention de l'entité et le rappel estimé de cette annotation est de 91%) par :

$$\text{Facilité}(e_i) = \frac{\#corrects}{\#corrects + \#incorrects}$$

Un score supérieur (resp. inférieur) à 0,5 signifie que lors de l'annotation un plus grand nombre de documents corrects qu'incorrects ont été trouvés (resp. plus d'incorrects que de corrects).

Ce graphique montre que les scores de nos approches ainsi que ceux des autres participants semblent, à quelques exceptions près, corrélés à la "facilité" de l'entité.

Pour la plupart des entités, nos approches obtiennent un score supérieur au score médian et cela semble d'autant plus vrai que les entités sont "difficiles" ($< 0,4$) (plus particulièrement *James McCartney*, *Vladimir Potanin*, *Lisa Bloom* et *Charlie Savage*). Un élément d'explication nous est apporté par la figure 4 qui présente le nombre de documents par entité et par heure. Deux de ces entités (*McCartney* et *Savage*) présentent des successions de pics importants et la prise en compte dans notre approche du facteur temps est probablement à l'origine de ces bons résultats. Cependant, la présence de successions de pics ne signifie pas systématiquement l'obtention de bons résultats comme le démontre le test avec l'entité *Nassim Nicholas Taleb*. Les résultats pour cette entité sont étonnamment bas alors que le topic devrait être plutôt facile de par l'importante quantité de documents pertinents. Les faibles résultats obtenus pour *Basic Element*, le groupe musical, et *Boris Berezovksy*, le pianiste, s'expliquent principalement par l'existence d'homonymes qui sont de plus très présents dans le corpus.

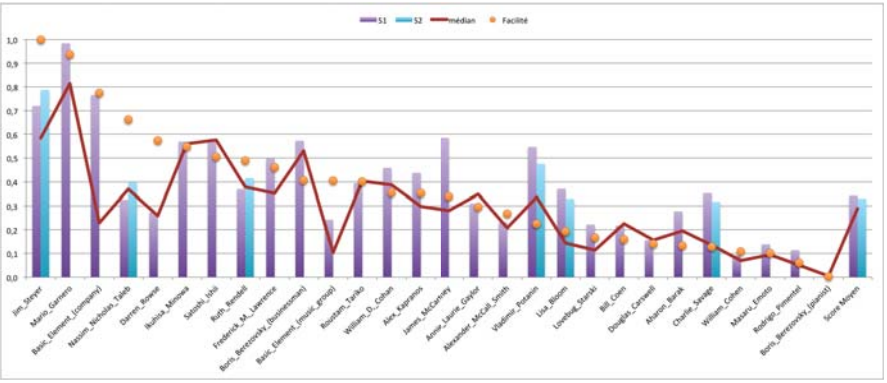

Figure 3. *Résultats officiels pour les documents "central" entité par entité pour la F-mesure. Sont représentés : la facilité d'une entité, le score médian ainsi que ceux de nos deux soumissions (pour une meilleure lisibilité, un seul résultat est affiché lorsque les deux soumissions se valent).*

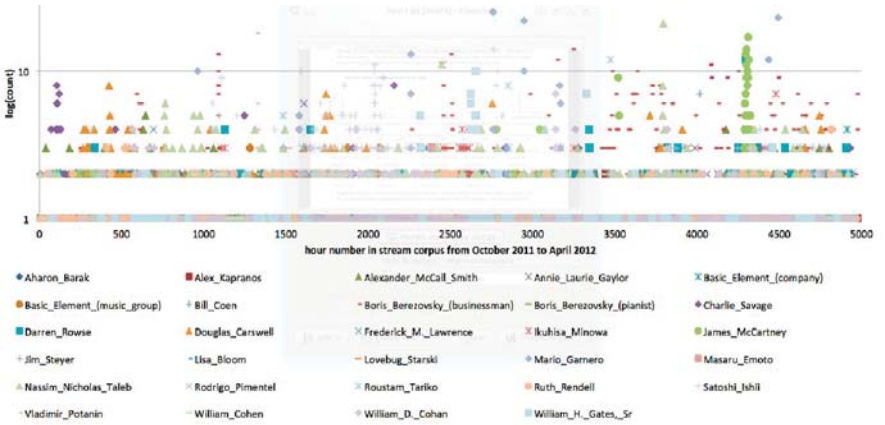

Figure 4. *Nombre de documents par entité et par heure pour l'intégralité du corpus.*

Nos deux soumissions obtiennent des résultats très proches, aux écarts presque toujours non significatifs. Pourtant, des différences existent : l'approche pour laquelle les documents classés comme non pertinents ne sont pas exclus (utilisant la fonction de score S2), obtient des résultats légèrement supérieurs pour les topics faciles et, à l'inverse, des résultats inférieurs pour les entités difficiles. Cela fait sens car par entité facile nous entendons une entité pour laquelle les documents la mentionnant sont plus souvent jugés pertinents que l'inverse. Ainsi, le fait de filtrer les documents jugés non

pertinents par le classifieur se traduit par une forte baisse du rappel alors que c'est un avantage pour les entités difficiles car le gain en précision est plus important.

Par ailleurs, la figure 5 aide à distinguer les caractéristiques des documents centraux par rapport aux autres : elle présente les proportions de documents (du corpus d'apprentissage) "intéressants" (en rouge) et "centraux" (en bleu) pour les différentes valeurs des critères utilisés. Dans ce corpus, les proportions de documents "intéressants" et "centraux" sont de deux pour un (voir en bas à droite de la figure). L'intuition liée à chaque critère devrait se traduire sur les histogrammes par une seule classe associée aux documents pour les valeurs extrêmes : "intéressante" pour les valeurs faibles et "central" pour les hautes. Cependant, un critère à lui seul ne peut dans un problème si complexe, séparer parfaitement les classes. La répartition des documents de chaque classe en fonction des valeurs pour un critère donné, indique son pouvoir discriminant : si la proportion de documents associés à la classe "central" (resp. "intéressant") pour les valeurs élevées (resp. faibles) est plus importante qu'un tiers (resp. deux tiers) alors le critère permet de caractériser dans une certaine mesure les documents de cette classe (voir par exemple "Nbr_Docs_Avec_Mentions_24H"). À l'inverse, un critère pour lequel les proportions de documents de chaque classe sont équivalentes au deux pour un et quelle que soit la valeur ne permet pas de prendre des décisions (par exemple "Nbr_Mentions_0_10P_Doc").

Pour la plupart des critères, les documents avec des valeurs élevées sont majoritairement de la classe "central" et cela conforte donc nos choix de critères. Entre autres, la présence d'un nombre important d'occurrences dans un document ainsi que la présence d'entités liées semble caractériser fortement l'intérêt de celui-ci. Les critères caractérisant la période d'apparition des documents tels que le nombre de mentions de l'entité dans les documents les 7 derniers jours ou le nombre moyen de documents contenant une mention semblent aussi très importants.

Au contraire, et ceci est surprenant, la présence ou non d'une mention de l'entité dans le titre semble avoir une corrélation très faible voire nulle avec l'intérêt du document étudié. De plus le nombre d'occurrences dans le titre des documents des 7 derniers jours fait au contraire partie des critères les plus discriminants.

La position des occurrences dans le document semble ne pas avoir de lien non plus avec la classe d'un document alors qu'en résumé automatique une approche basique mais performante consiste à considérer les phrases en début et fin de document comme particulièrement importantes. Une première explication est que si ce résultat est vérifié pour des documents journalistiques, il est peut-être moins transposable à des entrées de blogs ou de forums qui constituent les deux tiers de la collection KBA.

Enfin, pour certains critères, comme le nombre de documents faisant mention d'une entité, en 24 heures, ou encore l'écart type de cette valeur sur 7 jours, on remarquera qu'un groupe de documents peu pertinents ont des valeurs très élevées alors même qu'il semblait y avoir une corrélation entre des valeurs élevées pour ces critères et une forte proportion de documents importants. Nous pensons que la raison est l'absence de normalisation pour les valeurs de ces critères entre entités (une valeur

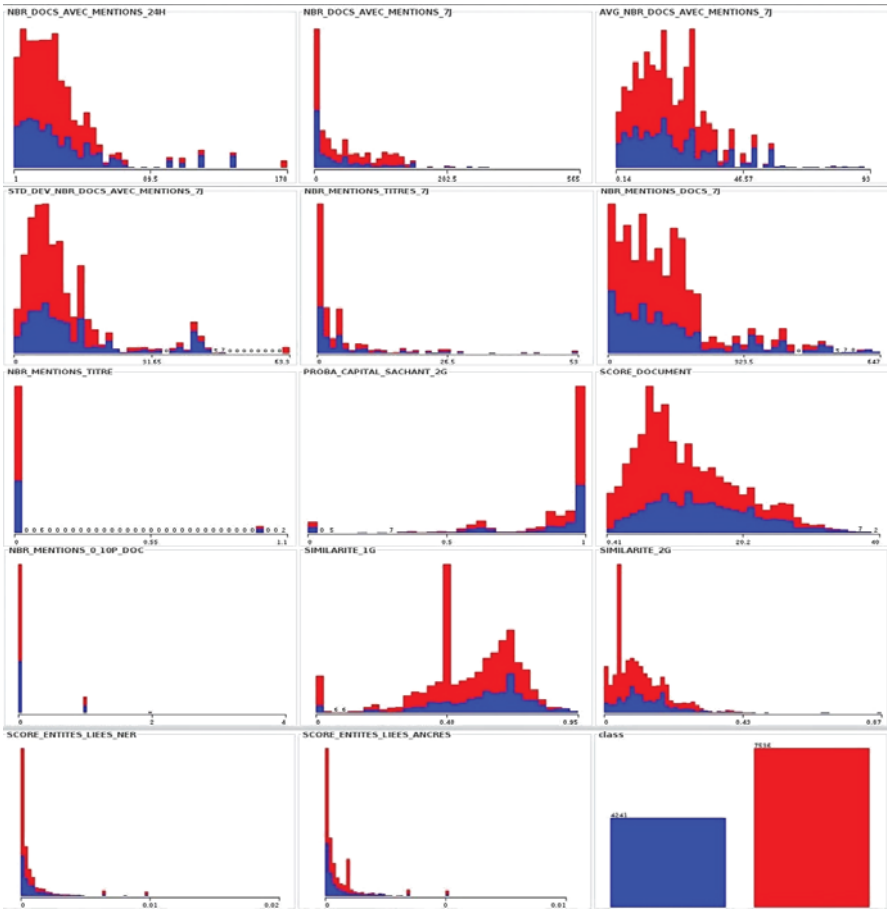

Figure 5. *Proportions de documents "intéressants" ou "centraux" en fonction des valeurs des différents critères utilisés pour l'apprentissage de* $c_{i,c}$

élevée n'a pas la même signification pour une entité couramment citée et pour une dont l'actualité est plus rare). En effet, une grande majorité des documents ayant ce comportement inhabituel sont associés à l'entité Boris Berezovsky qui est à la fois la plus citée sur le corpus (près de 2500 fois contre 500 fois en moyenne par entité) et avec le plus grand nombre de pics (voir figure 4).

L'un des avantages des forêts d'arbres décisionnels est qu'il est possible de mesurer l'utilité de chaque critère dans la séparation des classes dans les arbres (via la mesure de pureté de Gini). Ces résultats ne sont pas présentés ici par soucis de place mais surtout car ils sont en parfaite adéquation avec l'analyse qui vient d'être faite.

4. Conclusions et perspectives

Nous avons présenté une méthode supervisée pour estimer la degré d'intérêt d'un nouveau document dans un flux au regard d'une entité présente dans une base de connaissances. Cette approche est basée sur des critères tels que le nombre de mentions dans le document, la présence ou non d'entités liées ou encore différents indices mesurant l'ampleur de l'actualité de l'entité dans la période d'apparition du nouveau document. La pertinence de notre approche a été validée par de bons résultats obtenus dans le cadre d'une participation à la tâche KBA de la campagne d'évaluation TREC 2012. Elle nous a permis de mesurer la justesse de nos intuitions dans le choix des critères utilisés même s'il semble que des approches plus simples puissent s'avérer plus performantes. Cela souligne l'intérêt et la difficulté de la tâche mais aussi la marge de progression, importante pour le futur.

L'un des principaux résultats et que certains critères, largement répandus en recherche d'information et résumé automatique, se sont révélés beaucoup moins discriminants qu'escompté (la position des occurrences des entités dans les documents, la présence ou non de l'entité dans le titre...) même si leur pouvoir discriminant pourrait être amélioré en normalisant les valeurs entre entités.

Enfin, nous pensons à la nécessité de mettre à jour l'entrée de l'entité dans la base au cours du temps. Cela permettrait entre autres d'opérer des connexions avec les problématiques liées à la population automatique de bases de connaissances, notamment dans le cadre de la campagne d'évaluation "Knowledge Base Population" à TAC.

5. Bibliographie

Araujo S., Gebremeskel G., He J., Bosscarino C., de Vries A., « CWI at TREC 2012, KBA Track and Session Trac », *Proceedings of The 21th TREC*, 2012.

Artiles J., Li Q., Cassidy T., Tamang S., Ji H., « CUNY BLENDER TAC-KBP2011 Temporal Slot Filling System Description », *Proceedings of the Fourth TAC*, 2011.

Asahara M., Matsumoto Y., « Japanese Named Entity Extraction with Redundant Morphological Analysis. », *Proc. of the Human Language Technology conference - ACL*, 2003.

Atkinson J., Bull V., « A multi-strategy approach to biological named entity recognition », *Expert Systems with Applications, Vol. 39, No. 17*, 2012.

Bikel D., Miller S., Schwartz R., Weischedel R., « Nymble : a High-Performance Learning Name-finder », *Proc. Conference on Applied Natural Language Processing*, 1997.

Bonnefoy L., Bouvier V., Bellot P., « LSIS/LIA at TREC 2012 Knowledge Base Acceleration », *Proceedings of The 21th TREC*, 2012.

Breiman L., « Random Forests », *Machine Learning, Vol. 45 No. 1*, 2001.

Cancedda N., Goutte C., Renders J.-M., Cesa-Bianchi N., Conconi A., Li Y., Shawe-Taylor J., Vinokourov A., Graepel T., Gentile C., « Kernel Methods for Document Filtering », *Proceedings of The 11th TREC*, 2002.

Clark J., Gonzalez-Brenes J., « Coreference Resolution : Current Trends and Future Directions », 2008.

Collins-Thompson K., Ogilvie P., Zhang Y., Callan J., « Information Filtering, Novelty Detection, and Named-Page Finding », *Proceedings of The 11th TREC*, 2002.

Cucchiarelli A., Velardi P., « Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence », *Computational Linguistics 27 :1.123-131*, 2001.

Cucerzan S., « Large-scale named entity disambiguation based on Wikipedia data », *Proceedings of the 2007 Joint Conference on EMNLP and CoNLL*, 2007.

Das D., Martins A., « A Survey on Automatic Text Summarization », 2007.

Davis A., Veloso A., da Silva A., Jr. W. M., Laender A., « Named Entity Disambiguation in Streaming Data », *Proceedings of the 50th Annual Meeting of the ACL*, 2012.

Frank J., Kleiman-Weiner M., Roberts D., Niu F., Zhang C., Ré C., « Building an Entity-Centric Stream Filtering Test Collection for TREC 2012 », *Proceedings of The 21th TREC*, 2012.

Ji H., Grishman R., Dang H., « Overview of the TAC2011 Knowledge Base Population Track », *Proceedings of the Fourth TAC*, 2011.

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

Lee J., « Mining spatio-temporal information on microblogging streams using a density-based online clustering method », *Expert Systems with Applications, Vol. 39 No. 10*, 2012.

Liu X., Wei F., Zhang S., Zhou M., « Named Entity Recognition for Tweets », *ACM Transactions on Intelligent Systems and Technology, Vol. 9, No. 4*, 2012.

McCallum A., « Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-Enhanced Lexicons », *Proc. Conference on Computational Natural Language Learning*, 2003.

Mihalcea R., « Classifier Stacking and Voting for Text Filtering », *Proceedings of The 11th TREC*, 2002.

Nadeau D., Sekine S., « A survey of named entity recognition and classification », *Linguisticae Investigationes, Vol. 30, No. 1*, 2007.

Pasca M., Lin D., Bigham J., Lifchits A., Jain A., « Organizing and Searching the World Wide Web of Facts-Step One : The One-Million Fact Extraction Challenge », *Proc. National Conference on Artificial Intelligence*, 2006.

Qi X., Davison B., « Web page classification : Features and algorithms », *ACM Computing Surveys (CSUR) Vol. 41n No. 2*, 2009.

Robertson S., Soboroff I., « The TREC 2002 Filtering Track Report », *Proceedings of The 11th TREC*, 2002a.

Robertson S., Walker S., Zaragoza H., Herbrich R., « Microsoft Cambridge at TREC 2002 : Filtering Track », *Proceedings of The 11th TREC*, 2002b.

Sekine S., « Nyu : Description of the Japanese NE System Used For Met-2 », *Proc. Message Understanding Conference.*, 1998.

Sekine S., Nobata C., « Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy », *Proc. Conference on Language Resources and Evaluation*, 2004.

Voorhees E. M., « The TREC-8 Question Answering Track Report », *NIST Special Publication 500-246 : The Eighth Text REtrieval Conference (TREC-8)*, 1999.