Une approche non supervisée pour le typage et la validation d'une réponse à une question en langage naturel : application à la tâche *Entity* de TREC 2010

Ludovic Bonnefoy[*,**], **Patrice Bellot**[*], **Michel Benoit**[**]

**Université d'Avignon - CERI/LIA*
Agroparc – B.P. 1228
84911 Avignon Cedex 9 (France)
patrice.bellot@univ-avignon.fr

***iSmart*
Le Mercure A
13851 Aix-en-Provence Cedex 3 (France)
{ludovic.bonnefoy,michel.benoit}@ismart.fr

RÉSUMÉ. *La recherche d'entités nommées a été le sujet de nombreux travaux en recherche d'information. Dans ce papier, nous cherchons à déterminer si une entité est d'un type donné, et ce de manière non-supervisée et quel que soit son type. Nous proposons pour cela une approche basée sur l'utilisation de modèles de langage estimés à partir du web. De plus, nous souhaitons déterminer si cette nouvelle information peut être utilisée efficacement pour améliorer le classement des réponses (entités) candidates à une question en langue naturelle. Afin d'évaluer ces deux points, nous avons participé à la tâche Entity à TREC 2010.*

ABSTRACT. *Searching for named entities has been the subject of many researches in information retrieval. In this paper, we seek to determine whether a named entity is of a given type and in what extent it is. We propose to address this issue by an unsupervised web oriented language modeling approach. In addition, we want to determine if this new information can be used to improve the ranking of candidate answers (entities) in respect to a natural language question. To evaluate these two points, we participated to the TREC 2010 Entity track.*

MOTS-CLÉS : *questions-réponses, TREC Entity, divergence de Kullback-Leibler, typage d'entités nommées*

KEYWORDS: *question answering, TREC Entity, Kullback-Leibler divergence , named entities identification*

1. Introduction

1.1. *La reconnaissance des entités nommées*

Depuis une quinzaine d'années les entités nommées sont au cœur de nombreux travaux dans le domaine de la recherche d'information ou du traitement de la langue naturelle écrite en général (résumé automatique, ontologies, ...). Ce développement est en partie dû au fait que de multiples campagnes d'évaluation ont accordé une part importante à leur recherche ou utilisation au sein de leurs pistes tels que MUC (*Named Entity task*[1]), TREC (avec la tâche *Question Answering* (Voorhees, 1999)), ...

Les premières méthodes de recherche d'entités nommées, en l'absence de corpus d'apprentissage, se basaient 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) de nombreuses méthodes d'apprentissage automatique 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 dites semi-supervisées ont aussi été étudiées telle le *bootstrapping* qui consiste à démarrer d'un petit nombre d'exemples et de l'agrandir par itérations successives en exploitant différents critères comme les relations syntaxiques (Cucchiarelli *et al.,* 2001) ou synonymiques (Pasca *et al.,* 2006).

La reconnaissance des entités nommées est centrale dans bon nombre de problématiques en recherche d'information comme par exemple Questions-Réponses (QR). Cette tâche a connu un fort engouement ces dernières années. En effet, on a pu voir plusieurs campagnes d'évaluation internationales en faire un sujet important (TREC, CLEF, INEX, Equer, ...). Un système de QR présente au moins deux différences par rapport à un système de RI. La première est la formulation de la requête qui est une phrase interrogative en langage naturel (par exemple *"Je veux connaître les spécifications techniques du nouveau blackberry"*). Cela a de l'intérêt pour les utilisateurs (la formulation de requêtes efficaces sous forme de mots clés est une tâche difficile) et pour les systèmes (apport d'un contexte et d'informations supplémentaires). La seconde principale différence est la forme des résultats : un moteur de RI va retourner une liste de documents, dans lesquels l'utilisateur va être en charge de trouver la réponse par lui même, tandis qu'en QR, le système doit retourner une série de réponses précises (c'est-à-dire des chaînes correspondant exactement à ce que l'utilisateur recherche), généralement des entités nommées. C'est pourquoi une identification correcte (localisation et typage) des entités est une étape vitale pour de tels systèmes. Dans cet article, nous proposons une méthode non-supervisée permettant de déterminer à quel point une entité (dans son sens le plus large) est d'un type donné.

Cette méthode devrait permettre de traiter de manière intéressante un sujet encore peu étudié qui est de déterminer la proximité ontologique et sémantique de deux en-

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

tités et, au-delà, dans quelle mesure deux entités sont similaires. Cette problématique est intéressante comme l'atteste la création en 2009 de la piste *Entity Relation Finding* à TREC, sa poursuite en 2010 et sa planification pour 2011[2].

Le dernier point, et probablement le plus important, est d'arriver à traiter ce problème pour n'importe quel type d'entités et pas seulement les quelques types de très haut niveau (personnes, lieux, organisation, dates,...) que l'on a l'habitude de rencontrer depuis les campagnes MUC (Nadeau *et al.,* 2007) ou les quelques dizaines de types plus fins (c'est-à-dire des sous-catégories des types de haut niveau (Sekine *et al.,* 2002)) qu'exploitent certains systèmes. Notre objectif est de pouvoir traiter de manière égale et automatique des types aussi généraux que "personne" ou beaucoup plus fins, tels que "coéquipier" ou encore "distilleries de whisky".

1.2. *Entity track à TREC*

La tâche Entity proposée à TREC depuis 2009 se définit comme trouver une liste d'entités nommées associées à leur *homepage* sur le web à partir d'un topic[3] composé d'une entité nommée source, le type d'entités nommées attendues (parmi quatre types de haut niveau : *personne*, *lieu*, *organisation* et *produit*) et une partie *narrative* exprimant en texte libre (phrase ou question) la relation qu'elles doivent satisfaire avec l'entité source. Les *homepages* doivent être trouvées dans le corpus ClueWeb09[4] qui contient 500 millions de pages web en anglais. De plus malgré les quatre types très généraux qui sont donnés, il est possible d'extraire du texte libre un type attendu de plus bas niveau (donc plus fin) (des sous-types des quatre pré-cités tels que *compagnie aérienne*, *président*, ...) nous permettant d'évaluer le dernier point évoqué.

Cet article est composé de la manière suivante : dans une première partie, nous présentons une solution pour mesurer le degré d'appartenance d'une entité nommée à un type donné de manière non-supervisée. Dans une seconde partie, nous présentons la solution mise en œuvre pour répondre au sujet de la tâche Entity à TREC et comment nous y avons inclus notre proposition. Dans une troisième partie, nous analysons et commentons les résultats obtenus lors de cette campagne d'évaluation et enfin dans une dernière partie nous concluons et avançons quelques perspectives.

---

2. http ://ilps.science.uva.nl/trec-entity/2010/11/plans-for-entity-2011/
3. Exemple de topic :

```
<query>
   <num>21</num>
   <entity_name>Bethesda, Maryland</entity_name>
   <entity_URL>clueweb09-en0004-43-35557</entity_URL>
   <target_entity>location</target_entity>
   <narrative>What art galleries are located in Bethesda, Maryland ?</narrative>
</query>
```

4. http ://boston.lti.cs.cmu.edu/Data/clueweb09/

2. Mesure de l'appartenance d'une entité à un type donné

Comme nous l'avons évoqué plus haut, nous aspirons ici à une méthode efficace pour déterminer si une entité est d'un type donné, et ce, sans apprentissage au préalable, afin de se passer de corpus coûteux et limitant le nombre de types que l'on peut traiter. C'est pourquoi nous avons opté pour une approche orientée web.

Nous avons commencé par analyser manuellement les pages web retournées par des moteurs de recherche commerciaux en les interrogeant au sujet de quelques types d'entités généraux. Nous nous sommes très rapidement aperçu que les pages associées à chaque type avaient tendance à posséder un vocabulaire spécifique. Par exemple, pour le type *"portable mp3 players"*, certains mots comme *"mp3", "music", "capacity", "headphones",...* ont une fréquence d'occurrence nettement supérieure à celle qu'ils ont dans un corpus générique (c'est-à-dire tout ensemble très large de documents, traitant de toutes sortes de sujets).

En analysant ensuite les pages web associées à des entités en particulier, nous avons vérifié que, pour chacune d'entre elles, l'on obtenait des probabilités d'apparition des termes généralement éloignées de celles que l'on trouve dans un corpus générique (par exemple pour *Winnie l'ourson* certains mot comme *"fictional", "character", "bear", "friends", "disney", ...* ont une probabilité d'apparition élevée).

Notre dernière observation est que la distribution des probabilités d'apparition des termes, dans les pages associées à une entité donnée, est proche de celle des termes dans les pages web associées au *type* la caractérisant le plus (par exemple, pour un *"iPod"* certains mots comme *"apple","mp3","music","headphones","media", ...* ont une fréquence élevée tout comme pour l'ensemble des pages répondant à la requête *"portable mp3 players"*).

L'idée de la méthode que nous avons mise en œuvre découle de ces observations. Elle consiste à comparer le modèle de langage L_E (c'est-à-dire la distribution de probabilité des termes dans une collection) associé à une entité donnée à un modèle de langage générique L_{type} associé à un type d'entité donné.

Les étapes sont les suivantes :

1) collecter un ensemble de pages web liées à une entité (ex : *"Isaac Asimov"*), et un second ensemble, appelé *"ensemble de référence"*, correspondant au type de l'entité (ex : *"science-fiction writers"*). Ces pages sont, en ce qui nous concerne, récupérées en interrogeant le web via un moteur de recherche commercial ;

2) estimer, pour chaque ensemble, leur modèle de langage correspondant, où la probabilité d'apparition des termes (ici des unigrammes) est lissée avec Dirichlet :

$$p'(w|E) = \begin{cases} p_s(w|E) \text{ si w est présent dans l'ensemble } E \\ \alpha_d p(w|C) \text{ sinon} \end{cases} \quad [1]$$

où $p'(w|E)$ est la probabilité du mot w dans l'ensemble de pages web E, $p_s(w|E)$ est la probabilité lissée de w, $p(w|C)$ est la probabilité du mot w dans une collection C (consiste ici en 10% du corpus ClueWeb09, choisis aléatoirement) lissée avec Laplace

et α_s est un multiplieur. $p_s(w|E)$ et α_s sont estimés de la manière suivante :

$$p_s(w|E) = \frac{tf(w,E) + \mu.p(w|C)}{\sum_{w' \in V} tf(w',E) + \mu} \qquad \alpha_d = \frac{\mu}{\sum_{w \in V} tf(w,E) + \mu} \qquad [2]$$

où $tf(w,E)$ est le nombre d'occurrences du mot w dans l'ensemble E, V est l'ensemble des mots w' présents dans E et μ est un facteur dont la valeur est empiriquement fixée à 2000 (valeur choisie par (Chen *et al.,* 1998) pour de larges collections journalistiques).

3) comparer les probabilités p'_Ed'apparition des termes dans les documents associés à l'entité à celles de *référence* p'_{type} associées au type. Pour cela nous calculons la divergence de *Kullback-Leibler* (KLD) entre les deux modèles :

$$KLD(E, type) = \sum_i p'_E(i).log \frac{p'_E(i)}{p'_{type}(i)} \qquad [3]$$

où $KLD(E, type)$ est la divergence de Kullback-Leibler pour une entité E donnée et le type, $p'_E(i)$ (resp. $p'_{type}(i)$) est la probabilité d'apparition du i^e mot dans les documents associés à l'entité E (resp. au type $type$).

Cette méthode propose ainsi une manière de calculer le degré d'appartenance de n'importe quelle entité donnée à n'importe quel type donné. Maintenant que nous avons un moyen de calculer ce degré d'appartenance, nous nous proposons d'exploiter cette nouvelle information au sein d'un système de QR et de l'évaluer dans le cadre de la tâche *Entity* à TREC 2010.

3. Application : Related Entity Finding à TREC 2010

3.1. *Vue générale de la méthode mise en œuvre*

Pour cette première participation à la tâche, nous avons pris le parti de décomposer notre système à la manière classique d'un système de QR (une structure plus ou moins similaire a été adoptée par la plupart des participants en 2010[5]) au vu des nombreuses similitudes que les deux tâches ont entre elles. Le fonctionnement général est le suivant (voir figure 1) : dans une première phase le topic est analysé afin d'en extraire les éléments importants puis est ensuite parfois étendu ou réécrit (Hold *et al.,* 2010)).

La deuxième phase consiste à récupérer un ensemble de documents pertinents grâce aux éléments extraits du topic.

La phase suivante est la reconnaissance et l'extraction des entités nommées candidates au sein de ce jeu de documents. Pour cela de nombreuses équipes ont eu recours à des outils de reconnaissance d'entités nommées tels que le Stanford-NER[6] ou LBJ-

5. http ://ilps.science.uva.nl/trec-entity/files/trec2010/trec2010-entity-workshop.pdf
6. http ://nlp.stanford.edu/ner/index.shtml

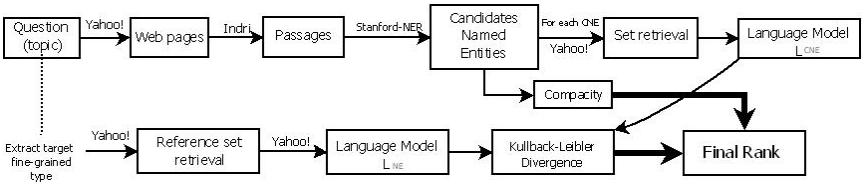

Figure 1. *Vue d'ensemble de Valen, notre système de typage et validation des réponses.*

based NER[7]. D'autres équipent ont étudié des voies différentes en 2010 et en 2009, comme l'utilisation des catégories de Wikipédia (en complément (Wang *et al.,* 2010b) ou non (Kaptein *et al.,* 2009a) d'un outil de reconnaissance) ou celles d'ontologies comme DBPedia, Yago ou encore Wordnet (Serdyukov *et al.,* 2009). D'autres manipulations sont ensuite parfois appliquées aux entités extraites afin d'en supprimer ou de les ramener vers une écriture canonique (Bron *et al.,* 2010).

L'étape suivante consiste à classer les entités nommées candidates. De nombreuses voies ont été explorées telles que l'estimation de la probabilité de générer une entité candidate donnée à partir d'un topic en utilisant le recouvrement des mots dans les documents supports et ceux du topic (Wu *et al.,* 2009) ou une similarité de type cosinus entre la *homepage* de l'entité candidate et le topic (Zhai *et al.,* 2009), le calcul de co-occurrences entre l'entité source et l'entité candidate (Bron *et al.,* 2009). De nombreux autres critères ont été utilisés, comme la distance en phrases entre l'entité candidate et l'entité source (Hold *et al.,* 2010), la fréquence de l'entité candidate (Wang *et al.,* 2010a) ou l'utilisation des liens hypertextes (Kaptein *et al.,* 2009b)

La dernière étape consiste à récupérer les *homepages* pour les entités candidates. Le recours à Wikipédia ou à d'autres bases de connaissance comme Freebase (Bron *et al.,* 2010) ou DBPedia (Wu *et al.,* 2010) a été choisi par certains participants à TREC 2010. D'autres ont travaillé sur les URLs, ou calculé de nombreux éléments et utilisé des approches d'apprentissage automatique (Hold *et al.,* 2010). (Wu *et al.,* 2010) montre que l'utilisation du score de confiance en la *homepage* trouvée pour re-classer les entités nommées candidates apporte une amélioration significative.

3.2. *Mise en œuvre détaillée pour notre participation à TREC 2010*

Comme cela a été dit plus haut, la première chose à faire pour pouvoir répondre à la tâche proposée est de récupérer un ensemble de pages web les plus pertinentes possible par rapport au topic. Pour cela nous avons choisi d'interroger le moteur Yahoo ! avec une requête constituée de l'entité source et des noms communs et noms propres de la partie *narrative* du topic (eux-mêmes reconnus via une analyse morpho-

7. http ://cogcomp.cs.illinois.edu/page/software_view/3

syntaxique faite avec l'étiqueteur TreeTagger[8]). Les 100 pages web les mieux classées sont récupérées, nettoyées des balises HTML et segmentées en passages d'une phrase (la longueur des passages semble n'avoir que peu d'impact sur les résultats (Gillard, 2006)). Les passages sont indexés avec Indri[9] qui est ensuite interrogé avec la requête précédemment élaborée et les 500 premiers passages sont conservés.

Nous effectuons une pré-sélection des entités nommées candidates en utilisant l'étiqueteur Stanford-NER pour les types "personne", "lieu" et "organisation". Pour les entités de type "produits", nous avons imaginé une autre méthode car ce type n'est pas présent dans les corpus de CoNLL et MUC sur lesquels le Stanford-NER est entraîné. Nous considérons, en premier lieu, comme produits candidats, toute séquence de noms propres ou de mots commençant par une lettre majuscule (ex : "Epson Stylus") et ne correspondant pas à une entité déjà détectée pour un autre type. De plus, si une de ces chaînes est immédiatement suivie par un nombre, nous le concaténons à cette séquence (ex : "Playstation 3").

Une normalisation des entités trouvées est effectuée pour les entités du type "personne" en associant les différentes écritures d'une entité nommée candidate (*Barichello*, *R. Barichello*, ...) vers leur forme canonique (*Rubens Barichello*) en conservant la forme la plus fréquente dans les extraits (*snippets*) proposés par Yahoo ! en lui soumettant comme requête chacune de ces écritures. De la sorte, on arrive à compléter des entités nommées où seul le nom de la personne serait récupéré (ex : "Rubens" $\rightarrow$ "Rubens Barichello") et réduire par deux ou trois la liste des candidats en supprimant de la redondance.

Le premier critère utilisé pour classer les entités nommées candidates est la compacité (Gillard *et al.,* 2006). La compacité mesure la proximité des mots de la requête autour d'une entité candidate (le postulat est que les réponses à une question tendent à apparaître à proximité des mots qui composent la question elle-même). La compacité est définie comme suit :

$$Compacité(E_i) = \frac{1}{|QW|} \sum_{w \in QW} \frac{Z_w}{|R_w| + 1} \quad [4]$$

avec QW l'ensemble des mots de la question (dans notre cas l'ensemble des éléments extraits du topic pour interroger les moteurs de recherche), $|QW|$ le cardinal de cet ensemble et w l'un de ces mots. Soit E_i une entité nommée candidate, R_w la distance (en nombre de mots) entre w et l'entité candidate. Soit Z le nombre de mots de la requête présents entre w et l'entité candidate E_i (inclus tous deux).

C'est en complément de cette métrique que nous utilisons la méthode présentée en section 2. Notre idée est de pénaliser d'autant plus une entité candidate qu'elle est éloignée (au sens de [3]) du type attendu d'entités. Cette approche est à mettre en relation avec les travaux de (Grappy *et al.,* 2010) qui, contrairement à nous, filtrent

8. http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
9. http ://www.lemurproject.org/indri/

les entités qui sont estimées comme n'étant pas du type souhaité et implémentent une approche basée sur un corpus d'apprentissage.

Dans la tâche *Entity Related Finding* à TREC 2010, seul un type d'entités nommées recherchées de haut niveau (très général) est explicitement donné. Cependant, nous avons remarqué qu'un type plus fin est exprimé dans la partie *narrative* de chaque topic. Pour un topic donné, nous avons choisi une stratégie d'extraction simple, et très *ad-hoc*, de cette information en sélectionnant le premier nom commun au pluriel dans le groupe nominal principal du champ *narrative* associé aux noms et adjectifs adjacents dans le cas où nous souhaitons un type plus précis encore.

Ensuite, pour chaque entité nommée candidate, nous calculons son degré d'appartenance au type fin que nous venons d'extraire. Un ensemble *de référence*, de 100 documents, va être récupéré pour le type en questionnant un moteur de recherche du web. Pour chaque entité candidate, nous constituons alors un ensemble de 10 pages. La divergence de Kullback-Leibler [3] va ensuite être calculée entre les modèles de langage estimés sur les différents ensembles de pages.

Nous avons maintenant à classer les entités nommées candidates en utilisant les différents scores à notre disposition. La première combinaison (nommée *Comp*), qui nous servira de *baseline*, consiste à n'utiliser que la compacité. La seconde approche (*Type*) consiste à n'utiliser que le degré d'appartenance [3] au type recherché de réponse. La troisième voie (*MH*) combine la compacité et l'appartenance. Elle consiste à associer à une réponse candidate la moyenne harmonique entre le rang attribué au regard de la compacité avec celui affecté pour son identification au type recherché.

Pour la dernière approche (*AA*), nous utilisons une méthode d'apprentissage automatique pour déterminer la meilleure manière de combiner différents scores. Pour chaque entité nommée candidate, nous cherchons à combiner efficacement quatre scores différents. Le premier est le meilleur score de compacité associé à cette entité nommée, le second est le degré d'appartenance au type, le troisième est l'idf ($idf(w_i) = log\frac{N}{n_i}$ avec N le nombre total de documents et n_i le nombre de documents où le mot w_i est présent) de l'entité dans les 500 premiers passages récupérés avec Indri et, enfin, le meilleur score attribué à un passage où l'entité nommée est présente. Les topics ont été choisis de manière à essayer de conserver les mêmes proportions pour chaque type ("personne", "lieu", ...) que dans le jeu de test de la tâche Entity à TREC 2009. Bien sûr, un ensemble d'apprentissage plus étendu mériterait d'être exploité. Ensuite, nous récupérons les entités nommées en sortie de notre système pour ces 45 topics et, pour chaque entité nommée en sortie, a été manuellement assignée une classe "Oui" si elle est réponse correcte au topic et "Non" sinon.

Avec cet ensemble de réponses correctes ou fausses, nous entraînons un classifieur de type perceptron multicouches (présent dans Weka[10])(permet d'obtenir les meilleurs résultats sur le jeu de test de TREC Entity 2009). Pour cette quatrième manière d'ordonner les réponses candidates, nous retenons comme fonction de score leur degré

10. www.cs.waikato.ac.nz/ml/weka/

d'appartenance à la classe apprise "Oui" (c'est-à-dire qu'une entité nommée pour laquelle le classifieur prédit que la réponse est correcte avec une confiance à 80% sera mieux classée qu'une autre pour laquelle le degré de confiance n'est que de 50%).

Maintenant que nous avons différentes méthodes pour classer les entités nommées candidates, il nous faut, pour répondre à la tâche de la campagne d'évaluation, rajouter en sortie du système de la figure 1 un module pour trouver une *homepage* correspondant à chaque entité nommée candidate dans le corpus ClueWeb09. La solution idéale ici aurait été d'indexer la partie anglaise du ClueWeb09. Cependant cette partie "pèse" près de 15 To et nous étions dans l'incapacité d'indexer une tel volume de données. Nous avons préféré faire la recherche de *homepages* sur le web pour ensuite trouver si la page est contenue dans le corpus[11] (stratégie par ailleurs choisie par la quasi-totalité des participants de TREC Entity 2010). Ayant constaté que la plupart des moteurs de recherche traitent de manière particulière le terme *homepage*, nous interrogeons le web à nouveau via le moteur Yahoo ! avec la requête *"Entité_Nommée homepage"* (par exemple *"Lufthansa homepage"*) et nous récupérons les cinq meilleures pages. Nous supprimons ensuite les pages web qui n'ont pas les caractéristiques d'une homepage. Pour cela nous avons entraîné un classifieur SVM sur le corpus *7-web genre*[12] (les genres sont, entre autres : *blog*, *shop*, *personnal homepage*, *frontpage*, ...) (les classifieurs SVMs sembleraient obtenir les meilleurs résultats pour cette tâche (Dewdney *et al.,* 2001)). Pour entraîner le classifieur nous utilisons de nombreuses caractéristiques comme la fréquence des mots (sans exceptions (Stamatatos *et al.,* 2000), la fréquence des POS (part-of-speech), le nombre de phrases, de mots, la taille moyenne en mots des documents, ... (Dewdney *et al.,* 2001) et le compte des marqueurs HTML (Levering *et al.,* 2008). Parmi les cinq pages récupérées, celles qui ne se voient pas attribuer la catégorie *homepage* sont supprimées. La page la mieux classée et présente dans le ClueWeb09 est associée à l'entité nommée candidate et constitue la sortie finale de notre système. Si aucune page n'a été trouvée alors l'entité n'est pas retenue.

4. Résultats numériques

Dans cette partie sont présentés les résultats officiels pour notre participation à la tâche Entity à TREC 2010. Cette évaluation était programmée pour être faite sur 50 topics créés pour l'occasion (numérotés de 21 à 70). Cependant trois d'entre eux ont été éliminés par les organisateurs car non pertinents. Quatre mesures ont été utilisées pour évaluer les approches : la précision à 10 éléments, MAP, nDCG@R et Rprec avec R le nombre de réponses correctes existantes. Lors de l'évaluation, l'accent a été mis par les organisateurs sur nDCG@R, qui favorise les systèmes apportant les meilleures réponses aux rangs les plus haut (en pénalisant les réponses correctes de manière lo-

11. Les organisateurs ayant fourni une correspondance des URLs vers les identifiants des pages dans le corpus
12. http ://www.webgenrewiki.org/index.php5/Genre_Collection_Repository

Métrique	Comp	Type	MH	AA	Meilleur	Médian
P@10	0,0468	0,0213	0,0362	**0,0532**		
nDCG@R	0,0737	0,0428	0,0610	**0,0766**	$\approx$ 0,39	0,0857
map	0,0261	0,0129	0,0200	**0,0305**		
Rprec	0,0463	0,0189	0,0373	**0,0591**		

Tableau 1. *Evaluation officielle de nos quatre approches pour la tâche Entity à TREC 2010 pour la précision à 10 éléments (P@10), nDCG@R, MAP et R-précision comparées au meilleur run et au run médian de tous les participants.*

garithmique en fonction de leur rang)[13], et c'est pourquoi c'est la seule d'entre elles pour laquelle il a été communiqué les scores du meilleur *run* et du médian.

Le tableau 1 présente les résultats que nous obtenons pour chaque méthode ainsi que le résultat du meilleur *run* de tous les participants à TREC 2010 et celui du *run* médian (si disponibles).

Ces résultats globaux permettent déjà de tirer des premières conclusions. La première observation est que pour notre première participation, notre meilleure approche se situe proche du score médian. Sur 48 runs, les nôtres occupent les positions 27, 29, 36 et 40. Il est à noter que parmi ces 48 runs, 19 d'entre eux sont en partie manuels (par exemple la sélection des mots clés à partir du topic). Si on ne se compare qu'avec des runs entièrement automatiques comme les nôtres, nos runs sont alors placés aux positions 14,16,18 et 21 (sur 29).

Si nous comparons nos différentes approches les unes aux autres, on peut tout d'abord constater, comme on pouvait s'y attendre, que n'utiliser que le degré d'appartenance d'une entité nommée candidate au type souhaité de réponse ne permet pas de correctement classer les entités entre elles. Cela est évidemment naturel car le contexte d'apparition de l'entité n'est pas pris en compte.

En second lieu, on peut s'étonner de la diminution du score obtenu par la baseline (compacité seule) lorsque l'on tente d'y ajouter l'information sur le type par le biais d'une moyenne harmonique des rangs. L'explication ici est que les deux informations ont de cette manière le même poids dans le classement. En observant les résultats topic par topic (tableau 4), on peut s'apercevoir que pour certains d'entre eux, le calcul du degré d'appartenance au type souhaité via des modèles de langage n'est pas adapté. En effet des types tels que *students* ou *winners* n'ont pas un modèle de langage assez spécifique (trop proches d'un modèle générique du monde) pour que la comparaison soit intéressante. De plus, ces types ne caractérisent pas de manière forte une entité (les pages concernant un étudiant ne vont pas vraiment mentionner ce fait mais plutôt les centres d'intérêts de cette personne). Il nous faudrait ici pénaliser le poids de cette information en fonction du degré de pertinence du type recherché (par exemple

13. http ://en.wikipedia.org/wiki/Discounted_cumulative_gain

Topic	Comp	Type	MH	AA	Meilleur	Médian
21	0,0166	**0,0213**	0,0177	0	0,4094	0,0260
22	0,0954	0,0852	0,096	**0,1009**	0,2818	0,1008
30	0,2342	0,1941	0,1705	**0,3155**	0,3739	0,0810
44	0	0	0	0	0,7099	0
49	0,139	0,1295	0,1321	**0,1451**	0,3081	0,1233
66	0	0	0	0	0,4306	0
67	**0,0443**	*0,0443*	0,0397	0,0260	0,4526	0,0725
Moyenne	0,0756	0,0678	0,0651	**0,0839**	0,4238	0,0577

Tableau 2. *Evaluation officielle pour le type "Lieu" de nos quatre approches pour la tâche Entity à TREC 2010 avec nDCG@R comparées au meilleur run et au run médian de tous les participants.*

Topic	Comp	Type	MH	AA	Meilleur	Médian
24	0,1873	0	0	**0,2743**	0,4348	0
37	0	0	0	0	0,6518	0
38	**0,3444**	0	0,2397	0,3046	0,6490	0,2193
41	**0,2979**	0,0289	0,2105	0,1940	0,4941	0,1614
43	**0,0232**	0,0502	0,0209	0,1440	0,6664	0,1172
52	**0,1586**	0,0221	0,1062	0	0,5829	0,0551
55	0	0	0	0	0,7661	0,0500
57	0	0	0	0	0,4693	0
Moyenne	**0,1264**	0,0127	0,0722	0,1146	0,5893	0,0754

Tableau 3. *Evaluation officielle pour le type "Personne" de nos quatre approches pour la tâche Entity à TREC 2010 avec nDCG@R comparées au meilleur run et au run médian de tous les participants.*

en fonction de la distance du modèle qui lui est associé avec un modèle de langage générique du monde).

Enfin, on peut voir que l'utilisation d'une méthode d'apprentissage automatique pour estimer le poids des différents scores pour le classement des entités nommées apporte des améliorations significatives : +14% pour P@10, 4% pour nDCG@R, 17% pour la MAP et +27% pour la R-précision. Cela vient donc confirmer notre intuition, montrant que l'utilisation du degré d'appartenance d'une entité nommée au type recherché permet d'améliorer les résultats obtenus par une métrique de type compacité.

Les résultats globaux sont intéressants mais ne permettent pas de tirer toutes les conclusions. Dans les tableaux 2, 3 et 4 sont présentés les résultats pour chaque type (excepté pour le type "produit" car un seul topic a finalement été conservé pour ce type, ce qui empêche toute analyse) au regard de la mesure nDCG@R.

La première chose importante que l'on remarque sur ces tableaux est que pour 10 des 47 topics plus de la moitié des participants obtiennent un score nul pour cette métrique mais, malgré tout, les meilleurs systèmes continuent d'obtenir de bons scores. Certains topics ont posé des problèmes pour la principale raison que les entités nommées réponses ne peuvent être trouvées dans la page Wikipédia de l'entité source (contrairement par exemple au topic 51). Dans notre cas, bien que nous ne traitons pas de manière particulière ces pages, elles figurent la plupart du temps dans les pre-

Topic	Comp	Type	MH	AA	Meilleur	Médian
23	0,1015	0	0,055	**0,1115**	0,4855	0,0550
25	0	0	0	**0,1365**	0,5718	0,0476
26	0,1223	**0,3273**	0,2821	0,2106	0,3998	0,0732
27	0	0	0	0	0,7638	0
29	**0,0087**	0	0,0073	0,0076	0,5216	0,2248
31	0,0383	**0,0730**	0,0562	0,0497	0,4627	0,0730
32	0	0	0	0	0,4522	0
33	0,2119	0,1798	**0,2275**	0,0601	0,3403	0,1111
34	0,0349	**0,045**	0,0381	0,0381	0,6573	0
36	0	0	0	0	0,3026	0
39	0,0504	0,0325	**0,0637**	0	0,5292	0,1404
40	0	0	0	0	0,5016	0,1361
42	0,0182	0,0138	**0,0246**	0,0108	0,3932	0,0682
45	**0,1203**	0	0	0,0907	0,6920	0,1203
47	0	0	0	0	0,7800	0,0913
48	0,08	0,0881	**0,0939**	0,0698	0,7628	0,1870
50	**0,0682**	0,0484	0,0595	0,0674	0,4626	0,1299
51	0,1713	0,1628	0,1596	**0,251**	0,5365	0,3807
53	0	0	0	0	0,5877	0
54	0,0488	**0,1198**	0,0337	0,0785	0,3175	0,1047
56	0	0	0	0	0,4171	0
58	0	0	0	0	0,3667	0
60	0	0	0	0	0,6988	0,0212
61	0,0373	0,0301	0,0184	**0,0502**	0,7115	0,0502
62	0	0	0	0	0,4715	0,1850
63	0,1466	0,1479	0,1361	**0,2004**	0,5255	0,1888
64	**0,2673**	0	0,255	0,0891	0,6814	0,1621
65	0	0	0	0	0,6131	0
68	0,0157	0,0404	0,0242	**0,0444**	0,4779	0,0978
69	0,3050	0,0428	0,1664	**0,4165**	0,7326	0,2880
70	0	0	0	0	0,5312	0
Moyenne	0,0596	0,0436	0,0549	**0,0640**	0,5403	0,0947

Tableau 4. *Evaluation officielle pour le type "Organisation" de nos quatre approches pour la tâche Entity à TREC 2010 avec nDCG@R comparées au meilleur run et au run médian de tous les participants.*

miers résultats des moteurs de recherche et font parties des pages dans lesquelles nous recherchons les entités nommées candidates. Cependant, certains équipes ont su exploiter d'autres ressources comme la *homepage* de l'entité source ou des bases de données (comme Freebase[14], DbPedia[15]...).

La seconde chose remarquable, c'est que pour deux des trois types (*personne* et *lieu*) nous obtenons des résultats supérieurs au *run* médian. On peut aussi voir que nos approches donnent de meilleurs résultats pour les *personnes* puis les *lieux* et enfin les *organisations*. L'ordre des performances suggère que nos résultats sont fortement impactés par ceux de l'étiqueteur Stanford-NER. En effet, on peut voir que l'ordre est directement corrélé à la capacité de reconnaissance des différents types (Finkel *et al.,* 2005). Il serait intéressant de mesurer dans un futur proche dans quelle mesure la pré-sélection des entités nommées candidates impact les résultats globaux et donc de déterminer son utilité. Enfin, la nette supériorité des résultats pour le type *personne*

14. http ://www.freebase.com/
15. http ://dbpedia.org/About

semble indiquer que notre méthode pour ramener les différentes écritures d'une même entité nommée vers la forme canonique est intéressante.

Le troisième élément important montré par ces résultats est le fait que l'utilisation de la divergence entre modèles de langage combinée de manière correcte avec la compacité apporte dans la plupart des cas une amélioration significative. Les topics du type *personne* font cependant exception et n'utiliser que la compacité semble pour ce type de questions préférable. Lorsque l'on regarde la capacité de la divergence seule à classer les entités nommées pour chacun des trois types on constate là aussi une grosse différence de comportement pour le type *personne*. En effet, pour les deux autres types, cette méthode, même si c'est celle qui obtient les plus bas scores, ne descend pas en dessous de 70% de celui de notre meilleur run. Pour le type *personne* en revanche, on chute à environ 10%. Cela confirme que cette approche n'est pas adaptée à ce type de question. La raison pour laquelle la divergence n'est pas adaptée ici a déjà été évoquée plus haut : la plupart des types fins extraits des topics ne sont pas assez spécifiques et ne caractérisent de manière forte les entités nommées (par exemple le type *members*) tandis que pour les deux autres types nous avons des classes nettement plus intéressantes (ex : *countries*, *airlines*, ...).

Enfin, le dernier point est la faible capacité de notre système à récupérer la *homepage* correspondante aux entités nommées candidates. Cette incapacité fait que le système est vraisemblablement amené à supprimer un grand nombre d'entités correctes (qui répondent au topic de manière juste) mais pour lesquelles nous n'avons pas réussi à identifier de *homepage*. L'exemple du tableau 6 illustre cela. La première colonne correspond aux cinq premières entités nommées candidates en sortie avant de rechercher pour chacune une *homepage*. La deuxième colonne montre les cinq premières entités nommées pour lesquelles a été trouvée une page à associer (peut-être n'est-ce pas vraiment une *homepage*). Les réponses en gras sont correctes et on peut voir que si la tâche ne consistait qu'à ramener des entités nommées alors sur ce topic on obtiendrait une précision à 5 éléments de 0,8 alors qu'en considérant l'identification des *homepages* comme sur les résultats officiels, nous n'avons qu'au mieux 0,2. Afin de confirmer cette piste, l'équipe de l'université de Potsdam nous a fourni son module de recherche des *homepages* (Hold *et al.,* 2010). Le tableau 6 présente les résultats obtenus avec leur module pour notre *run* avec la compacité uniquement et on peut y voir une très nette amélioration des résultats (de 22% à 87% selon la mesure). Ceci est clairement l'une des spécificités de la tâche que nous avons négligée et qu'il nous faudra améliorer lors de nos futures participations.

5. Conclusions et perspectives

Nous avons présenté une méthode automatique non-supervisée pour estimer dans quelle mesure une entité est d'un type donné (quel qu'il soit) basée sur la comparaison entre des modèles de langage associés à l'entité et au type. Notre participation à la piste Entity de la campagne d'évaluation TREC 2010 nous a permis de montrer que son utilisation peut améliorer les résultats des systèmes de question-réponse exploitant

Mesure	Valen + baseline *homepages*	Valen + HPFindingGoogle	Amélioration
P@10	0,0468	0,0574	+22%
nDCG@R	0,0737	0,0941	+28%
map	0,0261	0,0489	+87%
Rprec	0,0463	0,0724	+56%

Tableau 5. *Mesure de l'impact de la recherche des homepages sur les résultats. Comparaison des résultats obtenus pour Valen (notre système) avec notre baseline pour les homepages avec ceux obtenus avec Valen et le module de recherche des homepages de l'université de Potsdam pour les quatre mesures.*

Avant	Après
nico rosberg	nico rosberg
eddie irvine	**felipe massa**
felipe massa	muhammad ali
rubens barrichello	sebastian vettel
johnny herbert	joe louis

Tableau 6. *Les 5 premières entités candidates, avant et après la recherche des homepages, à partir du Topic 1 des questions d'entraînement de l'évaluation TREC 2009. Les réponses correctes sont en gras.*

par ailleurs des métriques de surface, telles que la *compacité*, en pénalisant les entités éloignées du type attendu. De plus, nous travaillons actuellement sur la constitution d'un corpus d'évaluation spécifique, afin d'estimer directement la qualité de notre méthode pour mesurer l'appartenance d'une entité à un type donné.

Les résultats obtenus lors de la campagne d'évaluation auraient pu être supérieurs si l'on avait mieux pris en compte certaines spécificités de la tâche. Il aurait notamment été intéressant de mettre au point une méthode plus sophistiquée pour trouver les *homepages* des entités nommées candidates et de prendre en compte, dans le reste du système, les informations présentes dans la *homepage* de l'entité source et dans sa page Wikipédia lorsqu'elle existe. Nous pensons aussi à la possibilité d'exploiter les informations contenues dans des bases de connaissances comme Freebase ou Wordnet pour déterminer le type d'une entité.

Retenons enfin que notre méthode de mesure d'appartenance d'une entité à un type d'entité possède des applications intéressantes pour l'aide à la constitutions de bases de connaissance puisqu'elle pourrait permettre d'identifier, à partir d'un exemple (instance) et du nom de sa catégorie (concept), d'autres instances similaires ou proches. L'étude des éléments linguistiques ayant permis cette identification pourrait à son tour conduire à définir les contextes d'apparition possibles des instances ainsi que certaines de leurs propriétés ontologiques. Tout cela fait l'objet de nos travaux actuels.

Remerciements

Merci à l'équipe de l'Université de Potsdam ayant participé à la tâche Entity de TREC 2010 de nous avoir permis d'utiliser leur module de recherche de *homepages*.

6. Bibliographie

Asahara M., Matsumoto Y., « Japanese Named Entity Extraction with Redundant Morphological Analysis. », *Proc. Human Language Technology conference - North American chapter of the Association for Computational Linguistics*, 2003.

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

Bron M., Balog K., de Rijke M., « Related Entity Finding Based on Co-Occurance », *NIST Special Publication 500-278 : TREC 2009*, 2009.

Bron M., He J., Hofmann K., Meij E., Tsagkias M., Weerkamp W., « The University of Amsterdam at TREC 2010 Session, Entity and Relevance Feedback », *The Nineteenth Text REtrieval Conference (TREC 2010) Notebook*, 2010.

Chen S. F., Goodman J., « An empirical study of smoothing techniques for language modeling. », 1998.

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

Dewdney N., VanEss-Dykema C., MacMillan R., « The form is the substance : classification of genres in text », *Annual Meeting of the ACL, Proceedings of the workshop on Human Language Technology and Knowledge Management - Volume 2001, Article 7*, 2001.

Finkel J. R., Grenager T., Manning C., « Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling », *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370*, 2005.

Gillard L., « ThÃ¨se », 2006.

Gillard L., Sitbon L., Blaudez E., Bellot P., El-Beze M., « Relevance Measures for Question Answering, The LIA at QA@CLEF-2006 », *Lecture Notes in Computer Science,4730/2007, Â« Evaluation of Multilingual and Multi-modal Information Retrieval Â», p. 440 Ã 449, 2007.*, 2006.

Grappy A., Grau B., « Validation du type de la réponse dans un système de questions réponses », *CORIA 2010, 7iÃ¨me Ã©dition de la ConfÃ©rence en Recherche d'Information et Applications*, 2010.

Hold A., Leben M., Emde B., Thiele C., Naumann F., « ECIR - A Lightweight Approach for Entity-Centric Information Retrieval », *The Nineteenth Text REtrieval Conference (TREC 2010) Notebook*, 2010.

Kaptein R., Koolen M., « Result Diversity and Entity Ranking Experiments : Anchors, Links, Text and Wikipedia », *Proceedings of The Eighteenth Text REtrieval Conference (TREC 2009)*, 2009a.

Kaptein R., Koolen M., Kamps J., « Result Diversity and Entity Ranking Experiments : Anchors, Links, Text and Wikipedia, University of Amsterdam », *NIST Special Publication 500-278 : TREC 2009*, 2009b.

Levering R., Cutler M., Yu L., « Visual Features for Fine-Grained Genre Classification of Web Pages », *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual 2008, pp. 131 - 131*, 2008.

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.

Nadeau D., Sekine S., « A survey of named entity recognition and classification », *Linguisticae Investigationes, Vol. 30, No. 1. (January 2007), pp. 3-26.*, 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.

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.

Sekine S., Sudo K., Nobata C., « Extended Named Entity Hierarchy », *Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC'02)*, 2002.

Serdyukov P., de Vries A., « Delft University at the TREC 2009 Entity Track : Ranking Wikipedia Entities », *Proceedings of The Eighteenth Text REtrieval Conference (TREC 2009)*, 2009.

Stamatatos E., Falotakis N., Kokkinakis G., « Text genre detection using common word frequencies », *Proceedings of the 18th conference on Computational linguistics - Vol. 2 (2000), pp. 808-814*, 2000.

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

Wang D., Wu Q., Chen H., Niu J., « A Multiple-Stage Framework for Related Entity Finding : FDWIM at TREC 2010 Entity Track », *The Nineteenth Text REtrieval Conference (TREC 2010) Notebook*, 2010a.

Wang Z., Tang C., Sun X., Ouyang H., « PRIS at TREC 2010 : Related Entity Finding Task of Entity Track », *The Nineteenth Text REtrieval Conference (TREC 2010) Notebook*, 2010b.

Wu Y., Kashioka H., « NiCT at TREC 2009 : Employing Three Models for Entity Ranking Track », *NIST Special Publication 500-278 : TREC 2009*, 2009.

Wu Y., Kawai H., « NiCT at TREC 2010 : Related Entity Finding », *The Nineteenth Text REtrieval Conference (TREC 2010) Notebook*, 2010.

Zhai H., Cheng X., Guo J., Xu H., Liu Y., « A Novel Framework for Related Entities Finding : ICTNET at TREC 2009 Entity Track », *NIST Special Publication 500-278 : TREC*, 2009.