Utilisation des relations d'une base de connaissances pour la désambiguïsation d'entités nommées

Romaric Besançon Hani Daher Olivier Ferret Hervé Le Borgne
CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, Gif-sur-Yvette, F-91191, France.
{romaric.besancon,hani.daher,olivier.ferret,herve.le-borgne}@cea.fr

RÉSUMÉ

L'identification des entités nommées dans un texte est une tâche essentielle des outils d'extraction d'information dans de nombreuses applications. Cette identification passe par la reconnaissance d'une mention d'entité dans le texte, ce qui a été très largement étudié, et par l'association des entités reconnues à des entités connues, présentes dans une base de connaissances. Cette association repose souvent sur une mesure de similarité entre le contexte textuel de la mention de l'entité et un contexte textuel de description des entités de la base de connaissances. Or, ce contexte de description n'est en général pas présent pour toutes les entités. Nous proposons d'exploiter les relations de la base de connaissances pour ajouter un indice de désambiguïsation pour ces entités. Nous évaluons notre travail sur des corpus d'évaluation standards en anglais issus de la tâche de désambiguïsation d'entités de la campagne TAC-KBP.

ABSTRACT

Using the Relations of a Knowledge Base to Improve Entity Linking

The identification of named entities in texts is an important task in Information Extraction tools, for numerous applications. This identification uses two steps : the recognition of named entity mentions in the text, that has been largely studied, and the association of the recognized mentions with known entities, appearing in a target Knowledge Base (Entity Linking). This second step usually relies on a similarity between the textual context of the entity mention and a description text associated with entities of the KB. However, in large KBs, this description text does not exists for all entities. We propose to exploit the relations present in the KB to add a disambiguation feature for the identification of the entities. We evaluate our approach on standard evaluation benchmarks for Entity Linking in English from the TAC-KBP evaluation campaign.

MOTS-CLÉS : Entités nommées, désambiguïsation, base de connaissances.

KEYWORDS: Named entities, entity linking, knowledge base.

1 Introduction

L'Extraction d'Information a pour but d'analyser automatiquement des textes pour en extraire une information structurée. Porté par plusieurs grandes campagnes d'évaluation telles que MUC (*Message Understanding Conference*), ACE (*Automatic Content Extraction*) et plus récemment TAC (*Text Analysis Conference*), ce domaine s'est structuré autour de tâches comme la reconnaissance automatique d'entités nommées dans les textes, l'identification de la coréférence, l'extraction de relations entre entités, ou l'extraction d'événements (qui peut être vue comme une extraction de

relations n-aires entre les entités), aussi appelé remplissage automatique de formulaires (*Slot filling*) (Piskorski & Yangarber, 2013).

Une étape supplémentaire de l'extraction d'information, identifiée plus récemment, est de faire automatiquement le lien entre les entités reconnues dans le texte et des entités connues, présentes dans une base de connaissances existante. Cette étape est appelée *désambiguïsation d'entités nommées* ou *Entity Linking*. Cette tâche s'inscrit parfois dans le cadre plus général d'une désambiguïsation globale de tous les concepts d'un texte par rapport à une base de connaissances, qu'ils soient portés par une entité nommée ou par une expression nominale : c'est le cas de systèmes comme Wikify (Mihalcea & Csomai, 2007) pour Wikipédia ou Babelfy (Moro *et al.*, 2014) pour la ressource encyclopédique BabelNet (Navigli & Ponzetto, 2012).

Les systèmes de désambiguïsation d'entités (Shen *et al.*, 2015; Ling *et al.*, 2015) s'appuient généralement sur deux types d'éléments : d'une part, des caractéristiques hors-contexte des entités, comme la similarité entre la chaîne de caractères de l'entité à désambiguïser et les formes connues des entités (noms, alias) dans la base de connaissances ; d'autre part, des caractéristiques contextuelles évaluant la similarité du contexte textuel de l'entité à désambiguïser avec un contexte textuel associé à l'entité cible dans la base de connaissances. Ce dernier est souvent un texte descriptif extrait de la page Wikipédia de l'entité. Or, dans des bases de connaissances de très grande taille comme Freebase, qui contient plusieurs dizaines de millions d'entités, nombre de ces entités n'ont aucune information de description ou de contexte textuel associée. De plus, les bases de connaissances sont en général structurées par des relations transverses qui lient les entités entre elles et qui forment une source d'information supplémentaire importante qui n'est, souvent, pas exploitée.

Nous proposons dans cet article d'exploiter cette information pour enrichir la notion de contexte des entités en prenant en compte leurs relations avec d'autres entités dans la base de connaissances, ce qui permet de pallier partiellement le manque de description textuelle associée aux entités. Dans notre approche, l'exploitation des relations se fait de façon simple, par l'utilisation d'une représentation vectorielle des relations permettant une mise en œuvre efficace de cette similarité, à la différence d'approches exploitant des structures sémantiques plus riches mais plus coûteuses comme AMR (*Abstract Meaning Representation*), utilisé par (Pan *et al.*, 2015). Nous évaluons notre approche sur des collections de référence annotées, en anglais, extraites de la tâche de peuplement de base de connaissances (KBP – *Knowledge Base Population*) des campagnes d'évaluation TAC, et plus particulièrement, de la tâche de découverte et désambiguïsation des entités (EDL – *Entity Discovery and Linking*), même si nous n'évaluerons que la partie désambiguïsation, en supposant que la reconnaissance des entités a déjà été effectuée.

Nous présentons l'approche proposée dans la section suivante, avant d'en proposer une évaluation dans la section 3, en détaillant les bases de connaissances et les collections utilisées, ainsi que les résultats obtenus, qui montrent que cette approche peut amener des améliorations sur la précision d'identification des entités, même si les résultats globaux restent similaires.

2 Description de l'approche

L'objectif de notre étude est d'évaluer l'impact de l'exploitation des relations présentes dans la base de connaissances sur la désambiguïsation des entités nommées. Pour cela, nous nous appuyons sur une approche pour la désambiguïsation qui traite de façon indépendante chaque entité du texte, après

que les entités ont été reconnues par un système indépendant de reconnaissance d'entités nommées. Des approches plus complexes ont été proposées, essayant par exemple d'effectuer de façon conjointe les tâches de reconnaissance et désambiguïsation des entités (Stern *et al.*, 2012; Luo *et al.*, 2015), en faisant une désambiguïsation jointe de toutes les entités du texte (Durrett & Klein, 2014; Chen & Ji, 2011) ou en combinant plusieurs modèles de désambiguïsation d'entités en agrégeant leurs résultats (Rizzo & Troncy, 2012; Ruiz *et al.*, 2015). Nous souhaitons ici vérifier l'intérêt du contexte relationnel pour la désambiguïsation des entités dans un cadre simple : les améliorations apportées ici pourront par la suite être intégrées dans ces systèmes plus complexes.

La tâche que nous considérons est donc, étant donné une mention d'entité dans un texte, de déterminer l'entité de la base de connaissance à laquelle elle réfère ou si elle n'est liée à aucune entité connue (on parle d'entité *NIL* pour la référence). Notre approche de désambiguïsation suit alors un processus standard (Ji *et al.*, 2014) : pour chaque mention d'entité à désambiguïser, les trois étapes suivantes sont ainsi réalisées :

1. l'analyse de la mention d'entité et de son contexte textuel ;
2. la génération d'entités candidates à partir de la base de connaissances ;
3. la sélection de la meilleure entité parmi les entités candidates.

Ces trois étapes sont présentées plus en détail dans les sections suivantes.

2.1 Analyse de la mention d'entité

Comme nous étudions la désambiguïsation des mentions d'entités nommées et non leur reconnaissance, nous nous plaçons dans un contexte où les mentions d'entités à désambiguïser sont données. Une étape de reconnaissance d'entités nommées dans le texte est néanmoins effectuée de façon complémentaire à cette entrée (nous avons utilisé l'outil MITIE [1]). Elle permet d'ajouter une information de type aux mentions d'entités à désambiguïser mais aussi de définir leur contexte en termes d'entités environnantes. En revanche, nous ne considérons que les mentions explicites d'entités nommées, en ignorant les mentions nominales ou pronominales. Nous n'avons pas, en effet, d'analyse de coréférence dans le processus d'analyse du document, ce qui serait nécessaire pour étendre la désambiguïsation à toutes les mentions d'entités. Dans l'analyse de la mention d'entité, le contexte textuel est utilisé à la fois pour enrichir la mention d'entité et pour calculer une représentation de son contexte.

Plus précisément, deux formes d'expansion de mention sont effectuées, qui peuvent être considérées comme des formes simples de coréférence :

— si la mention de l'entité est un acronyme, on cherche dans le document des entités nommées de même type que celle de l'entité acronyme, dont les initiales correspondent à l'acronyme. Ces entités sont alors ajoutées comme des variations de la mention d'entité ;
— si d'autres entités dans le document ont une expression incluant celle de la mention d'entité cible, elles sont ajoutées comme variations de la mention d'entité ;

Pour la représentation du contexte, une représentation vectorielle de type tf-idf est construite. L'espace vectoriel support de cette représentation est construit à partir de l'ensemble des documents Wikipédia associés aux entités de la base de connaissances.

1. https://github.com/mit-nlp/MITIE

2.2 Génération des candidats

La génération des entités candidates est effectuée en comparant la mention d'entité et chacune de ses variations obtenues lors de l'expansion de la mention d'entité avec les entités de la base de connaissances en utilisant les similarités suivantes (Dredze *et al.*, 2010) :

1. égalité des chaînes de caractères entre la mention et la forme normalisée de l'entité dans la base de connaissances ;
2. égalité des chaînes de caractères entre la mention et une variation (alias ou traduction) d'une entité de la base de connaissances ;
3. similarité de chaînes de caractères avec une variation d'une entité de la base de connaissances fondée sur une distance d'édition (distance de Levenshtein ≤ 2). Pour une meilleure efficacité, nous avons exploité une structure de BK-tree (Burkhard & Keller, 1973) pour cette fonctionnalité ;
4. inclusion de l'entité dans la chaîne de caractères d'une variation d'une entité de la base de connaissances (fonctionnalité implémentée par la base de données utilisée).

2.3 Sélection d'un candidat

2.3.1 Mesures de similarité

Pour sélectionner la meilleure entité cible parmi les entités candidates, on s'appuie sur deux scores de similarité. Le premier score mesure, de façon standard, la similarité entre le contexte textuel de la mention d'entité et le contexte textuel de l'entité de la base de connaissances. Plus précisément, cette similarité est mesurée par la distance cosinus entre les représentations vectorielles du contexte de la mention d'entité et du texte associé à l'entité candidate (*i.e.* le texte de sa page Wikipédia).

Le second score de similarité exploite les relations entre les entités dans la base de connaissances. Plus précisément, on veut déterminer si les entités qui apparaissent dans le contexte textuel de la mention d'entité sont proches des entités en relation avec l'entité candidate dans la base de connaissances. Dans toute sa complexité, cette approche demanderait de désambiguïser les autres entités du texte (ce qui reviendrait, en pratique, à faire une désambiguïsation jointe de toutes les entités) et de réaliser une extraction explicite de relations au niveau du texte, afin de vérifier que la cooccurrence d'une autre entité à proximité de la mention d'entité correspond explicitement à l'expression d'une relation entre ces entités. Pour ne pas traiter ces différents problèmes, ce qui risquerait d'entraîner un bruit important en cumulant les erreurs apportées par chacun des traitements, nous adoptons une approche plus simple en définissant une notion de *contexte relationnel* associé, comme le contexte textuel, à une représentation vectorielle. Pour chaque entité E de la base de connaissances, nous construisons une liste des entités en relation (binaire ou n-aire) avec E dans la base de connaissances. L'énumération des noms des entités de cette liste forme un pseudo-document sur lequel on applique le même processus de représentation vectorielle que pour le contexte textuel, dans le même espace défini par l'analyse des textes de la base de connaissances. Le score de similarité des relations est alors estimé par la distance cosinus entre ce vecteur et le vecteur représentant le contexte textuel de la mention d'entité.

Des exemples de contextes relationnels sont proposés par la figure 1 pour deux entités ambiguës de la base de connaissances Freebase ayant pour nom Texas (l'état des États-Unis et le groupe de musique).

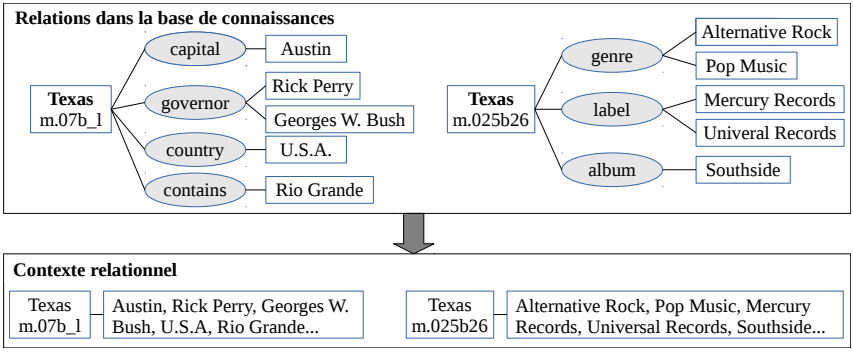

FIGURE 1 – Exemple de contexte relationnel pour deux entités ambiguës

2.3.2 Processus de sélection

Pour intégrer les différents critères associés à une entité candidate de manière flexible et choisir les informations les plus pertinentes de façon automatique, nous nous appuyons sur un système de classification statistique. Chaque entité candidate est représentée par un ensemble de traits comprenant les deux mesures de similarité présentées à la section précédente ainsi que quatre caractéristiques binaires indiquant l'origine de la génération du candidat (points 1 à 4 dans la section 2.2). Ces traits sont ajoutés avec l'idée qu'une égalité ou une forte similarité de forme est un bon indicateur de la qualité du candidat, indépendamment de la similarité du contexte.

Un classifieur est donc construit pour reconnaître la meilleure entité parmi les candidats grâce à un corpus d'entraînement. Plus précisément, nous utilisons un classifieur binaire décidant, pour une liste de caractéristiques données correspondant à une paire <mention d'entité, entité candidate>, si la mention est une instance de l'entité candidate. Les entités candidates sont générées pour toutes les entités mentions des données d'entraînement : les exemples positifs sont les instances des entités de référence parmi les candidats ; les exemples négatifs sont les autres entités candidates générées. Comme le nombre d'entités candidates générées pour chaque mention d'entité peut être très élevé (selon les collections, entre 1 et plusieurs centaines de milliers), nous limitons le nombre d'exemples négatifs à X fois le nombre d'exemples positifs, avec $X = 10$ dans toutes les expériences présentées.

Pour chaque mention d'entité, le classifieur associe une probabilité à chaque candidat, le candidat avec la plus forte probabilité étant alors sélectionné.

2.3.3 Filtrage par le type d'entité

Selon les collections de référence et les bases de connaissances, certaines tâches de désambiguïsation des entités sont associées à une tâche de typage des entités, en ne considérant que certains types d'entités. Dans la campagne TAC-EDL 2015, les types d'entités considérés sont ainsi : personne (PER), lieu (LOC), organisation (ORG), entité géopolitique (GPE), établissement (FAC). Pour ne pas dépendre de la qualité de l'outil de reconnaissance des entités nommées appliqué lors de l'analyse de la mention d'entité, nous n'utilisons pas le type d'entité produit par cet outil et générons les entités candidates indépendamment de leur type.

Un filtrage selon le type est alors effectué durant le processus de sélection du meilleur candidat. Plus précisément, nous gardons les 5 meilleurs résultats retournés par le classifieur, filtrons les candidats dont le type, donné par la base de connaissances, est incompatible avec un des types attendus, et sélectionnons le meilleur candidat parmi les candidats restants, s'il y en a. Pour la base de connaissances Freebase, des règles complémentaires ont été définies pour mesurer cette compatibilité : par exemple *division administrative* (*administrative_division*) ou *pays* (*country*) sont des types d'entités de Freebase compatibles avec le type GPE.

Une mention d'entité peut donc être marquée comme n'ayant pas d'entité de référence (NIL) si (1) aucune entité candidate n'est trouvée lors de la génération des candidats (2) tous les candidats sont rejetés par le classifieur (3) les candidats conservés sont rejetés par le filtrage sur le type d'entité.

3 Évaluation

3.1 Mesures d'évaluation

Pour les mesures d'évaluation, nous faisons appel aux mesures standards de précision, rappel et F-score pour la bonne reconnaissance de l'entité de référence lorsque celle-ci existe (*link*), la bonne reconnaissance d'une mention d'entité sans entité de référence (*nil*) et les résultats combinés (*all*). Par rapport aux références des campagnes d'évaluation TAC, la bonne reconnaissance du type des entités ainsi que le regroupement des entités NIL n'est pas pris en compte. Les mesures considérées correspondent aux mesures officielles *strong_link_match*, *strong_nil_match*, *strong_all_match* de la campagne TAC-KBP 2015. Lors des campagnes d'évaluation TAB-KBP précédentes, les scores utilisés étaient la précision $préc.(all)$, nommée *overall accuracy* et le rappel $rap.(link)$, nommé *KB accuracy*.

Plus précisément, si, pour une mention d'entité e, on note e_{ref} l'entité qui lui est associée dans la référence et e_{test} l'entité qui lui est associée de façon automatique par notre système de désambiguïsation, et $N(x)$ le nombre d'entités mentions qui vérifient la contrainte x, alors les mesures d'évaluation sont définies par les formules suivantes :

$$préc.(nil) = \frac{N(e_{test} = \text{NIL} \wedge e_{ref} = \text{NIL})}{N(e_{test} = \text{NIL})} \quad rap.(nil) = \frac{N(e_{test} = \text{NIL} \wedge e_{ref} = \text{NIL})}{N(e_{ref} = \text{NIL})}$$

$$préc.(link) = \frac{N(e_{test} = e_{ref} \wedge e_{test} \neq \text{NIL})}{N(e_{test} \neq \text{NIL})} \quad rap.(link) = \frac{N(e_{test} = e_{ref} \wedge e_{test} \neq \text{NIL})}{N(e_{ref} \neq \text{NIL})}$$

$$préc.(all) = \frac{N(e_{test} = e_{ref})}{N(e_{test})}$$

3.2 Collections et bases de connaissances

Nous évaluons notre approche avec plusieurs collections et deux bases de connaissances différentes : DBPedia [2] (Lehmann *et al.*, 2015), utilisée comme référence lors des campagnes d'évaluation TAC-

2. http://dbpedia.org

KBP de 2009 à 2013, est une extraction formalisée des informations de Wikipédia ; Freebase [3] (Bollacker *et al.*, 2008), qui a été au centre de la campagne d'évaluation TAC-KBP 2015, est une base de connaissances plus importante extraite de nombreuses sources et pouvant être enrichie par ses utilisateurs.

Nous présentons dans le tableau 1 des statistiques sur la taille des collections utilisées pour cette évaluation, en indiquant le nombre de documents et d'entités considérés. Pour les collections de TAC 2009 à 2013, l'évaluation était guidée par les entités : la tâche était de désambiguïser une entité particulière dans un document, ce qui explique un nombre d'entités proche du nombre de documents. Dans la collection 2015, la tâche est de désambiguïser toutes les entités d'un document, ce qui justifie l'équilibre très différent entre les nombres de documents et d'entités : il y a en moyenne près de 77 entités par document (les collections sont formées de dépêches de presse et d'extraits de forums).

	DBPedia			Freebase	
	Nb docs	Nb entités		Nb docs	Nb entités
TAC 2009	3688	3 904	TAC 2015/training	168	12 175
TAC 2010	2231	2 250	TAC 2015/test	167	13 587
TAC 2011	2231	2 250			
TAC 2012	2016	2 226			
TAC 2013	1820	2 190			

TABLE 1 – Taille des collections pour l'évaluation de la désambiguïsation des entités

En ce qui concerne les bases de connaissances, nous utilisons dans les deux cas une représentation commune sous la forme d'une base de données relationnelle contenant les informations suivantes : les entités, les variations des entités (les alias, formes dérivées, autres noms des entités ainsi que leur traduction dans différentes langues), une table pour les relations binaires entre deux entités, une table pour les propriétés des entités (les relations avec des valeurs littérales), une table pour les relations n-aires entre plusieurs entités ou avec des valeurs littérales (par exemple, un mariage est une relation n-aire entre 2 entités de type Personne et une propriété de type Date). Dans Freebase, ces relations n-aires sont appelées *Compound Value Type* (CVT).

Pour Freebase, un filtre est appliqué pour exclure les entités de certains types non considérés dans la campagne TAC-KBP 2015 (*book.written_work, book.book, music.release, music.album, tv.tv_series.episode music.composition music.recording, film.film and fictional_universe.fictional_character*). Nous présentons les statistiques des bases de connaissances dans le tableau 2, en indiquant le nombre d'entités, le nombre total de variations pour ces entités (qui sont un indicateur de l'ambiguïté à gérer) ainsi que le nombre d'entités avec des descriptions textuelles associées et celles avec des relations associées.

	Entités	Variations	Entités avec contexte	Entités avec relations
DBPedia	818 741	4 276 395	818 670	687 964
Freebase	8 707 140	21 744 088	4 659 531	5 683 818

TABLE 2 – Taille des bases de connaissances fondées sur DBPedia et Freebase

3. http://www.freebase.com

3.3 Résultats avec DBPedia

Nous présentons dans cette section les résultats des tests effectués en utilisant la base de connaissances DBPedia pour les références d'entités et les collections de documents des campagnes d'évaluation TAC-KBP des années 2009 à 2013. Nous donnons dans le tableau 3 quelques statistiques sur la génération des entités candidates pour ces collections, notamment le nombre d'entités mentions, le nombre d'entités NIL (les entités mentions pour lesquelles il n'y a pas de réponse attendue selon la référence), le nombre total de candidats générés, le nombre d'entités mentions pour lesquelles aucun candidat n'est généré, le nombre moyen de candidats générés et le rappel sur les candidats, défini comme le pourcentage d'entités mentions pour lesquelles la réponse attendue est présente dans la liste des candidats générés. On note que ce score de rappel est plutôt bon (87,4% en moyenne), pour une méthode de génération des candidats relativement simple, avec un nombre moyen de candidats générés également limité (le nombre maximum de candidats se situe entre 2 718 et 9 964, selon les collections).

	Entités	Entités NIL	Candidats	Cand=0	Cand. moy.	rappel(Cand)
2009	3 904	2 229	208 060	949	70,41	84,0%
2010	2 250	2 230	232 672	601	141,10	89,4%
2011	2 250	1 126	329 508	388	176,96	87,9%
2012	2 226	1 049	420 179	117	199,23	92,4%
2013	2 190	1 007	394 217	395	219,62	83,5%

TABLE 3 – Statistiques sur la génération de candidats pour les collections DBPedia (TAC 2009 à 2013)

	F-score(all)				
	2009	2010	2011	2012	2013
adaboost	**77,4%**	77,1%	71,9%	51,8%	73,8%
svm_rbf	74,5%	78,8%	67,8%	**55,0%**	72,3%
svm_linear	74,3%	**80,4%**	**72,6%**	50,4%	**74,1%**
random_forest	70,8%	71,1%	61,1%	49,7%	68,8%
	F-score(link)				
	2009	2010	2011	2012	2013
adaboost	**67,5%**	70,7%	62,0%	**43,2%**	70,2%
svm_rbf	59,4%	64,3%	44,7%	33,6%	58,1%
svm_linear	64,5%	**72,6%**	**62,6%**	43,0%	**70,4%**
random_forest	62,8%	63,6%	52,9%	41,8%	64,4%

TABLE 4 – Résultats de la désambiguïsation des entités sur les collections DBPedia, en testant sur une collection et en utilisant toutes les autres pour l'entraînement

La classification est calculée sur un petit nombre de traits (six), qui ont des comportements différents (quatre traits binaires, deux à valeurs réelles). Nous avons choisi de tester pour cette tâche les classifieurs suivants : Adaboost, Machines à Vecteurs de Support (SVM) avec noyau RBF ou linéaire

et *Random Forest*. Le tableau 4 présente les résultats obtenus par le système de désambiguïsation proposé avec ces classifieurs (nous avons utilisé l'implémentation de *scikit-learn* (Pedregosa *et al.*, 2011), sans réaliser d'optimisation particulière de leurs paramètres). Pour chaque corpus de test, l'entraînement des classifieurs est effectué sur les données des autres années.

Les résultats obtenus sont globalement bons, de l'ordre de 75% pour la désambiguïsation globale, sauf pour la collection 2012 qui semble plus difficile que les autres (c'est la collection pour laquelle le rappel sur les candidats est le plus élevé, mais c'est également celle dont les entités mentions semblent les plus ambiguës parce que le nombre total de candidats est également le plus élevé). Les différences dues au choix du classifieur ne sont pas évidentes, même si, en tendance générale, Adaboost et les SVM linéaires donnent les meilleurs résultats.

3.4 Résultats avec Freebase

Pour la collection associée à Freebase, le nombre d'entités mentions considérées et le nombre d'entités dans la base de connaissances sont beaucoup plus importants. Le tableau 5 présente des statistiques sur les candidats pour les corpus d'entraînement et de test. Pour le corpus de test, nous nous sommes restreints aux mentions d'entités nommées en ignorant les mentions nominales. Ces

	Entités	Entités NIL	Candidats	Cand=0	Cand. moy.	rappel(Cand)
training	12 175	3 215	5 844 592	1 282	458,08	76,0%
test	13 587	3 379	6 141 369	1 255	480,32	77,6%

TABLE 5 – Statistiques sur les candidats sur la collection TAC 2015

résultats montrent que, pour la même méthode de génération de candidats que pour les collections DBPedia, le nombre de candidats est beaucoup plus élevé (ce qui s'explique simplement par la taille de la base de connaissances), et le rappel sur les candidats est également plus faible.

Dans le tableau 6, nous présentons les résultats pour la désambiguïsation des entités.

	F-score(all)	F-score(link)	F-score(nil)
adaboost	**63,6%**	**63,6%**	**63,4%**
svm_rbf	51,8%	48,1%	61,2%
svm_linear	57,4%	58,3%	58,6%
random_forest	59,6%	59,0%	61,0%

TABLE 6 – Résultats de la désambiguïsation des entités sur la collection Freebase

Ces résultats sont globalement moins bons que pour les collections DBPedia mais la tâche est plus difficile étant donnée l'augmentation de la taille des données, que ce soit pour la base de connaissances ou le nombre d'entités à désambiguïser. On obtient néanmoins des scores généraux corrects de plus de 60% avec le classifieur Adaboost (l'écart avec les autres classifieurs est plus important que dans les collections DBPedia). Par rapport aux résultats de la campagne d'évaluation, la figure 2 présente le positionnement de cette méthode (*adaboost*) par rapport aux résultats des participants de la campagne

(Ji *et al.*, 2015)[4]. On observe que nos résultats sont plutôt corrects, se plaçant en 5ème position avec 54,0%, derrière un groupe assez homogène autour de 65-70%, ces meilleurs scores pouvant s'expliquer par un meilleur traitement des types des entités NIL et une meilleure couverture dans la génération des candidats (cf. section suivante).

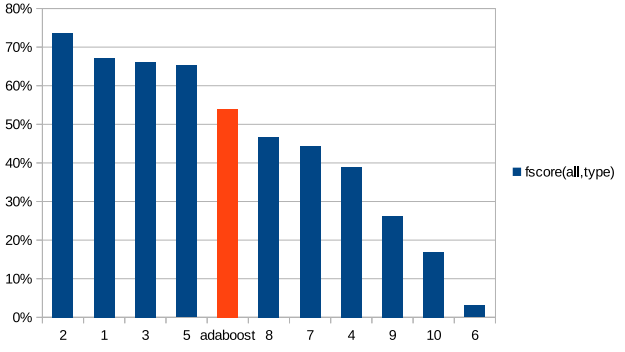

FIGURE 2 – Résultats obtenus avec le classifieur Adaboost comparés aux 10 soumissions des participants de la campagne TAC EDL 2015.

3.5 Analyse d'erreurs et discussion

3.5.1 Entités manquantes

Le tableau 7 présente le nombre d'entités attendues par la référence qui ne sont pas présentes parmi les candidats, à chaque étape du processus de désambiguïsation. La plupart des entités manquantes

Étape	nombre d'entités manquantes
Génération des candidats	2 287
Sélection des 5 meilleurs candidats	1 325
Filtrage sur le type	213
Sélection du meilleur candidat	967

TABLE 7 – Nombre d'entités attendues perdues à chaque étape du processus de désambiguïsation

sont absentes lors de la génération des candidates : une analyse plus précise de ces erreurs montre qu'un petit nombre (8) d'entités de la référence ne sont pas dans notre base de données (i.e. elles ont été filtrées lors de la construction de la base de connaissances à partir de la totalité de Freebase). Parmi les autres entités manquantes les plus fréquentes, on trouve des acronymes qui n'apparaissent pas explicitement dans leur forme étendue dans le contexte textuel de la mention et qui ne sont pas des variations d'entités présentes dans la base de connaissances (par exemple, l'acronyme *U.S.* n'est pas présent dans les variations de *United States of America* dans Freebase). Des étapes additionnelles de

4. Le score disponible pour tous les participants est un score de désambiguïsation prenant en compte le bon typage des entités, y compris pour les entités NIL : le fait de ne pas avoir traité ce problème explique que le score sur cette figure est plus bas que le score dans le tableau précédent.

mise en correspondance d'acronymes avec des entités ou d'enrichissement de la base de connaissances par des acronymes provenant d'autres sources ou générés automatiquement peuvent être envisagées pour pallier ce problème. Dans le même registre, les adjectifs de nationalité (*French*, *Chinese*, *American* etc.) ne sont pas associés aux pays dans Freebase (alors que ce lien est attendu dans la référence) : des ressources linguistiques additionnelles seraient nécessaires pour expliciter ce lien.

Les entités perdues durant la phase de filtrage sur le type d'entité indiquent soit que les règles de compatibilité entre les types d'entité de Freebase et les types attendus ne sont pas suffisantes, soit que des informations de type manquent dans Freebase.

Les méthodes de sélection des 5 meilleurs et du meilleur candidat les plus efficaces dépendent fortement, quant à elles, du type et des paramètres du classifieur utilisé : avec le meilleur classifieur (Adaboost), le nombre total d'entités correctes perdues dans ces deux étapes est diminué de 26% par rapport au résultat obtenu avec le classifieur *Random Forest*.

3.5.2 Impact de l'utilisation des relations

Nous présentons dans cette section une étude plus détaillée de l'impact de l'utilisation des relations de la base de connaissances sur le processus de désambiguïsation des entités. Pour Freebase, en particulier, les entités de la base de connaissances ne sont pas restreintes aux entités de Wikipédia et n'ont, par conséquent, pas toutes une page de description fournissant un contexte textuel à comparer avec le contexte de la mention d'entité à désambiguïser. Nous présentons dans le tableau 8 les résultats obtenus sur le corpus Freebase avec et sans le score de similarité sur le contexte relationnel, en utilisant le classifieur Adaboost, qui donne les meilleurs résultats globaux. Les résultats montrent

	avec les relations			sans les relations		
mesure	précis.	rappel	F-score	précis.	rappel	F-score
link	**69,6%**	59,0%	**63,8%**	58,8%	**69,6%**	63,7%
nil	53,4%	78,1%	63,4%	**54,5%**	**80,1%**	**64,9%**
all	63,7%	63,7%	63,7%	**64,1%**	**64,1%**	**64,1%**

TABLE 8 – Résultats sur le corpus Freebase, avec et sans la prise en compte des relations

que l'ajout de la similarité sur les relations tend à augmenter de façon significative la précision des entités non NIL, mais cette augmentation est contrebalancée par une baisse du rappel. Sur les entités NIL, les scores sont meilleurs sans les relations. Les scores globaux sont du même ordre (le classifieur Adaboost ayant une composante aléatoire, des scores moyens pourraient être calculés sur plusieurs tests pour avoir des comparaisons plus robustes).

Comme précisé précédemment, une des motivations pour exploiter les relations de la base de connaissances est de gérer les entités qui n'ont pas de description textuelle associée. Le tableau 9 présente le nombre d'entités correctes (selon la référence) qui ont une valeur de similarité non nulle sur le contexte textuel et sur les relations. On voit dans ce tableau que moins de 0,5% des entités ont une valeur de 0 pour les deux similarités. 4,5% des entités ont une valeur 0 pour la similarité du contexte textuel (soit parce que l'entité n'a pas de description associée, soit parce que l'intersection de sa description avec le document de l'entité est vide, sur le vocabulaire de représentation choisi) mais une valeur non nulle pour la similarité sur les relations. Parmi ces 357 entités, 129 sont gardées comme choix final par le classifieur : même si ces entités n'auraient pas pu être trouvées sans la similarité de

similarités	nb. entités
sim(doc)≠0 and sim(rel)≠0	6 967
sim(doc)≠0 and sim(rel)=0	580
sim(doc)=0 and sim(rel)≠0	357
sim(doc)=0 and sim(rel)=0	38

TABLE 9 – Nombre d'entités correctes selon leurs valeurs de similarité : sim(doc) est la similarité cosinus entre les contextes textuels et sim(rel) la similarité cosinus entre les contextes de relations

relations, elles représentent une portion réduite des entités mentions et n'ont pas un impact significatif sur le score global.

4 Conclusion

Nous présentons dans cet article une approche pour la désambiguïsation des entités qui exploite les relations de la base de connaissances pour aider la désambiguïsation. L'approche proposée repose sur une représentation simple de ces relations, sous la forme d'un vecteur de contexte relationnel, qui peut être comparé au vecteur de contexte textuel de l'entité à désambiguïser. Les premiers tests effectués sur des collections d'évaluation s'appuyant sur les bases de connaissances DBPedia et Freebase montrent des résultats encourageants : la précision de désambiguïsation des entités non NIL est augmentée notablement, même si les scores globaux restent comparables.

L'approche que nous avons proposée pour exploiter les relations de la base de connaissances est relativement simple. Nous envisageons de l'améliorer en prenant en compte plus d'information, à la fois du côté de la mention d'entité et du côté de la base de connaissances. En effet, nous considérons tous les mots du texte autour de la mention d'entité alors qu'en utilisant une composante d'extraction de relation dans l'analyse du document, on pourrait ne considérer que les entités du texte en relation explicite avec la mention d'entité. Puisque chaque entité en relation dans le document participe à la désambiguïsation des autres entités, une approche jointe (Durrett & Klein, 2014) ou collaborative (Chen & Ji, 2011) de la désambiguïsation des entités peut également être envisagée. D'un autre côté, nous ne considérons dans la base de connaissances que les relations directes avec l'entité candidate : ces relations pourraient être enrichies avec des relations de niveau 2 (i.e. ajouter les entités en relation avec les entités en relation directe). Néanmoins, une implémentation naïve de cette solution risque d'être coûteuse : en temps pour construire les vecteurs de représentation des entités de niveau 2 et en espace, les vecteurs construits étant moins creux. Une alternative pourrait être de construire une représentation dense des relations de la base de connaissances, ce qui permettrait de factoriser l'information des relations (Nickel *et al.*, 2012; Chang *et al.*, 2014).

Remerciements

Ces travaux sont partiellement financés par le projet F1409071Q CuratedMedia.

Références

BOLLACKER K., EVANS C., PARITOSH P., STURGE T. & TAYLOR J. (2008). Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, p. 1247–1250: ACM.

BURKHARD W. A. & KELLER R. M. (1973). Some Approaches to Best-match File Searching. *Communications of the ACM*, **16**(4), 230–236.

CHANG K.-W., TAU YIH W., YANG B. & MEEK C. (2014). Typed Tensor Decomposition of Knowledge Bases for Relation Extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*: Association for Computational Linguistics.

CHEN Z. & JI H. (2011). Collaborative Ranking: A Case Study on Entity Linking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, p. 771–781: Association for Computational Linguistics.

DREDZE M., MCNAMEE P., RAO D., GERBER A. & FININ T. (2010). Entity Disambiguation for Knowledge Base Population. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, p. 277–285: Association for Computational Linguistics.

DURRETT G. & KLEIN D. (2014). A Joint Model for Entity Analysis: Coreference, Typing, and Linking. *Transactions of the Association for Computational Linguistics*, **2**, 477–490.

JI H., NOTHMAN J. & HACHEY B. (2014). Overview of TAC-KBP2014 Entity Discovery and Linking Tasks. In *Text Analysis Conference (TAC)*.

JI H., NOTHMAN J., HACHEY B. & FLORIAN R. (2015). Overview of TAC-KBP2015 Tri-lingual Entity Discovery and Linking. In *Text Analysis Conference (TAC)*.

LEHMANN J., ISELE R., JAKOB M., JENTZSCH A., KONTOKOSTAS D., MENDES P., HELLMANN S., MORSEY M., VAN KLEEF P., AUER S. & BIZER C. (2015). DBpedia – A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, **6**(2), 167–195.

LING X., SINGH S. & WELD D. (2015). Design Challenges for Entity Linking. *Transactions of the Association for Computational Linguistics*, **3**, 315–328.

LUO G., HUANG X., LIN C.-Y. & NIE Z. (2015). Joint Entity Recognition and Disambiguation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 879–888, Lisbon, Portugal: Association for Computational Linguistics.

MIHALCEA R. & CSOMAI A. (2007). Wikify! Linking Documents to Encyclopedic Knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management (CIKM'07)*, p. 233–242, New York, NY, USA: ACM.

MORO A., RAGANATO A. & NAVIGLI R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, **2**, 231–244.

NAVIGLI R. & PONZETTO S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, **193**, 217–250.

NICKEL M., TRESP V. & KRIEGEL H.-P. (2012). Factorizing YAGO: Scalable Machine Learning for Linked Data. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)*, p. 271–280: ACM.

PAN X., CASSIDY T., HERMJAKOB U., JI H. & KNIGHT K. (2015). Unsupervised Entity Linking with Abstract Meaning Representation. In R. MIHALCEA, J. Y. CHAI & A. SARKAR, Eds., *HLT-NAACL*, p. 1130–1139: Association for Computational Linguistics.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PISKORSKI J. & YANGARBER R. (2013). Information Extraction: Past, Present and Future. In T. POIBEAU, H. SAGGION, J. PISKORSKI & R. YANGARBER, Eds., *Multi-source, Multilingual Information Extraction and Summarization*, p. 23–49. Berlin, Heidelberg: Springer Berlin Heidelberg.

RIZZO G. & TRONCY R. (2012). NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, p. 73–76, Stroudsburg, PA, USA: Association for Computational Linguistics.

RUIZ P., POIBEAU T. & MÉLANIE F. (2015). ELCO3: Entity Linking with Corpus Coherence Combining Open Source Annotators. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, p. 46–50, Denver, Colorado: Association for Computational Linguistics.

SHEN W., WANG J. & HAN J. (2015). Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, **27**(2), 443–460.

STERN R., SAGOT B. & BÉCHET F. (2012). A Joint Named Entity Recognition and Entity Linking System. In *EACL 2012 Workshop on Innovative hybrid approaches to the processing of textual data*, p. 52–60, Avignon, France.