Désambiguïsation d'entités nommées par apprentissage de modèles d'entités à large échelle

Hani Daher — Romaric Besançon — Olivier Ferret — Hervé Le Borgne — Anne-Laure Daquo — Youssef Tamaazousti

CEA, LIST, Laboratoire Vision et Ingénierie des Contenus
Gif-sur-Yvette, F-91191 France
{prenom.nom}@cea.fr

RÉSUMÉ. *La désambiguïsation d'entités consiste à lier automatiquement des mentions d'entités identifiées dans un texte et des entités présentes dans une base de connaissances. L'approche générale consiste à produire, pour une mention donnée, des entités candidates puis à sélectionner la meilleure parmi celles-ci, selon un ensemble de critères. Notre travail se focalise sur cette dernière étape, avec une méthode fondée sur l'apprentissage de modèles permettant d'opérer une discrimination entre une entité et les entités qui lui sont ambiguës. Un verrou majeur dans ce contexte réside dans la capacité à gérer de grandes bases de connaissances nécessitant l'apprentissage de dizaines de millions de modèles. Nous proposons trois stratégies permettant d'y répondre, offrant différents compromis entre efficacité et qualité de reconnaissance. Nous les validons expérimentalement sur six bases provenant des campagnes TAC en utilisant les bases de connaissances Freebase et DBpedia.*

ABSTRACT. *The objective of Entity Linking is to connect an entity mention in a text to a known entity in a knowledge base. The general approach for this task is to generate, for a given mention, a set of candidate entities from the base and determine, in a second step, the best one. This paper focuses on this last step and proposes a method based on learning a function that discriminates an entity from its most ambiguous ones. We adopt a model that is able to deal with large knowledge bases. Thus our contribution lies in the strategy to learn efficiently such a model. We propose three strategies with different efficiency/performance tradeoff. The approach is experimentally validated on six datasets of the TAC evaluation campaigns by using Freebase and DBpedia as reference knowledge bases.*

MOTS-CLÉS : *Recherche d'entités, Désambiguïsation d'entités, Sélection d'exemples négatifs.*

KEYWORDS: *Entity IR, Entity retrieval, Entity linking, Negative sample selection*

1 Introduction

La normalisation des entités nommées présentes dans les requêtes est connue pour son impact positif sur les processus de recherche d'information (Khalid *et al.*, 2008). Nous nous intéressons dans cet article à la tâche de désambiguïsation d'entités nommées (aussi appelée *Entity Linking*) consistant à faire automatiquement le lien entre des entités trouvées dans un texte et des entités connues, présentes dans une base de connaissances existante (Ling *et al.*, 2015 ; Shen *et al.*, 2015), aboutissant ainsi à une normalisation non équivoque desdites entités. Une telle tâche est parfois généralisée à un système plus complexe visant à désambiguïser globalement tous les concepts d'un texte par rapport à une base de connaissances donnée, que ce soient des entités nommées ou des expressions nominales (e.g. Wikify (Mihalcea et Csomai, 2007) ou Babelfy (Moro *et al.*, 2014)).

Un système de désambiguïsation d'entités comporte usuellement trois composants principaux (Ji *et al.*, 2014). Premièrement, le texte requête est analysé pour y identifier des « mentions d'entités » susceptibles d'être désambiguïsées au regard de la base de connaissances de référence. Ensuite, pour chaque mention d'entité, le système produit plusieurs « entités candidates » à partir de la base. Finalement, il sélectionne la meilleure entité parmi les candidates. L'une des principales difficultés, dans ce contexte, est de pouvoir gérer le très grand nombre d'entités généralement présentes dans la base.

La principale contribution de cet article concerne la dernière étape du processus de désambiguïsation : étant donné un ensemble d'entités candidates trouvées dans la base de connaissances pour une mention d'entité, nous proposons une nouvelle méthode de sélection de l'entité la plus susceptible d'être liée à cette mention. L'idée principale est de construire un modèle discriminant capable de distinguer une entité particulière en la différenciant des entités avec lesquelles elle partage le plus d'ambiguïté. Un tel modèle devant être appris pour chaque entité de la base et celle-ci en contenant plusieurs millions ou dizaines de millions, nous avons adopté un modèle linéaire dont la complexité reste compatible avec une telle taille, à la fois pendant l'apprentissage et durant la phase de test (susceptible d'être « en ligne »). Ainsi, le cœur de notre contribution est une méthode de construction de l'ensemble d'apprentissage de ces modèles et plus spécifiquement une méthode de sélection des exemples négatifs permettant d'apprendre plusieurs millions d'hyperplans.

Nous présentons l'état de l'art relatif à la désambiguïsation d'entités dans la partie 2. La partie 3 présente notre approche en détail. En particulier, le processus complet de notre système est décrit dans les parties 3.1 et 3.2, tandis que la stratégie d'apprentissage des modèles discriminants utilisés pour sélectionner la meilleure entité candidate est détaillée dans la partie 3.3. Finalement, nous évaluons notre système en utilisant les bases de référence Freebase et DBpedia et six collections de test provenant des campagnes d'évaluation TAC (*Text Analysis Conference*). Les résultats de cette évaluation sont présentés dans la partie 4.

2 État de l'art

Nous organisons la revue des méthodes de désambiguïsation d'entités selon leur degré de supervision : non supervisé, semi-supervisé et supervisé.

Les méthodes non supervisées s'appuient en général sur la simple définition d'un score de similarité entre la mention à désambiguïser et l'entité de la base de connaissances, la sélection s'effectuant sur la maximisation de ce score. Ces scores se fondent en général sur des recouvrements de contextes (Cucerzan, 2007) et peuvent combiner plusieurs mesures. Par exemple, (Han et Zhao, 2009) combine des similarités de mots et de concepts Wikipédia. Ces méthodes sont souvent simples et donc faciles à implémenter. Néanmoins, leurs performances sont aussi bien inférieures à celles des approches supervisées (Cassidy *et al.*, 2011) dans la plupart des cas.

Pour leur part, les méthodes supervisées sont généralement fondées sur des classifieurs binaires (Lehmann *et al.*, 2010 ; Varma *et al.*, 2009) ou des modèles d'ordonnancement (Shen *et al.*, 2012 ; Cao *et al.*, 2007) spécifiquement dédiés à la désambiguïsation des entités. Dans les deux cas, la problématique principale est liée à l'annotation des données d'entraînement, dont le coût devient prohibitif pour des bases contenant des millions d'entités telles que Freebase ou DBpedia. Parmi ces approches, certaines études (Zhang *et al.*, 2010 ; Fan *et al.*, 2015) utilisent des entités ambiguës pour apprendre des modèles identifiant l'entité correcte. Néanmoins, à notre connaissance, aucune approche ne propose d'apprendre un modèle discriminant par entité comme nous le faisons.

Plus précisément, (Zhang *et al.*, 2010) propose une méthode de construction d'exemples permettant d'apprendre un modèle de désambiguïsation. L'approche se concentre sur les mentions d'entités non ambiguës dans DBpedia. Le principe général est de générer des exemples de désambiguïsation en remplaçant dans les documents les mentions d'une entité qui ne sont pas ambiguës par des noms alternatifs de cette entité qui sont ambigus. Les exemples positifs sont construits en associant les documents modifiés avec l'entité tandis que les exemples négatifs sont produits en associant ces documents avec les entités faisant référence aux noms alternatifs. (Zhang *et al.*, 2011) utilise un algorithme d'apprentissage itératif pour sélectionner les entités les plus informatives qui sont proches de l'hyperplan séparateur.

(Fan *et al.*, 2015) utilise une stratégie de type « un contre tous » pour désambiguïser les entités. Considérant qu'il n'est pas possible d'apprendre un modèle pour chaque entité de Freebase, il propose une stratégie pour apprendre un seul classifieur global pour toutes les entités. Le principe est d'ajouter l'identifiant unique Freebase d'une entité comme caractéristique supplémentaire. Les exemples positifs d'une entité auront ainsi le même identifiant Freebase. Les exemples négatifs sont choisis aléatoirement parmi les entités ayant le même nom que l'entité. Pour le test, tous les identifiants des entités candidates (portant le même nom) sont testés par ce classifieur.

Enfin, dans le domaine des approches semi-supervisées, (Zheng *et al.*, 2012) se concentre sur le problème de la collecte des données et de leur annotation, plus pré-

cisément en exploitant à la fois un ensemble de données annotées et de données non annotées liées à Freebase. De façon itérative, un modèle est appris avec des exemples positifs construits à partir de phrases provenant d'articles Wikipédia contenant une forme non ambiguë d'une entité (constituant ainsi une vérité terrain fiable) et des exemples négatifs choisis aléatoirement parmi les autres entités. Le modèle appris permet d'annoter des documents non désambiguïsés, qui sont ensuite utilisés pour l'apprentissage lors de l'itération suivante.

3 Approche proposée

L'approche que nous proposons pour la désambiguïsation des mentions d'entités s'inscrit dans le paradigme de l'apprentissage supervisé. Cette désambiguïsation est ainsi réalisée par un classifieur entraîné préalablement en s'appuyant sur un ensemble d'exemples décrits par des traits. Nous proposons plus spécifiquement d'introduire un nouveau trait pour opérer cette classification, trait lui-même produit par un classifieur déterminant dans quelle mesure une mention d'entité correspond à une entité donnée de la base de connaissances de référence.

Plus globalement, notre système de désambiguïsation d'entités nommées adopte une architecture standard (Ji *et al.*, 2014) s'articulant autour de deux grandes étapes : sachant une mention d'entité et son contexte textuel, un premier module propose un ensemble d'entités candidates pour la désambiguïsation à partir de la base de connaissances de référence ; un second module évalue quant à lui les entités générées suivant un ensemble de critères, leur attribue un score pour caractériser cette évaluation et propose finalement l'entité de plus fort score comme référent de la mention d'entrée.

3.1 Génération des entités candidates

La génération des entités candidates pour une mention d'entité s'appuie à la fois sur une analyse intrinsèque de cette mention et une analyse de son contexte textuel. Dans cette étude, nous nous sommes focalisés principalement sur la désambiguïsation des entités et non sur leur identification. De ce fait, nous considérons les mentions d'entités à désambiguïser comme des données d'entrée. Une étape d'identification des entités nommées[1] est néanmoins réalisée afin d'associer un type (PERSONNE, LIEU, ORGANISATION) aux mentions d'entités et de définir leur contexte en termes d'entités environnantes (nous ne prenons en compte que les mentions explicites d'entités nommées en laissant de côté les mentions nominales et pronominales).

Deux formes d'expansion des mentions d'entités données, assimilables à des formes simples de coréférence, sont également menées :

1. Pour l'identification des entités nommées, nous utilisons l'outil MITIE : `https://github.com/mit-nlp/MITIE`

– si une mention d'entité est un acronyme, nous recherchons dans le document source les mentions d'entités de même type dont les initiales correspondent à l'acronyme ;

– nous recherchons les mentions d'entités du document source où la mention d'entité cible apparaît en tant que sous-chaîne.

Ces expansions sont ajoutées comme formes alternatives de la mention d'entité cible.

À la suite de cette phase d'analyse intrinsèque, des entités candidates sont générées en comparant la mention d'entité cible et ses formes alternatives[2] avec les entités présentes dans la base de connaissances de référence suivant quatre stratégies (Dredze *et al.*, 2010) :

– égalité entre la mention d'entité et une entité de la base de connaissances ;

– égalité entre la mention d'entité et une variation connue (alias ou traduction) d'une entité de la base de connaissances ;

– inclusion de la mention d'entité dans une des variations d'une entité de la base de connaissances ;

– similarité entre la mention d'entité et une variation d'une entité de la base de connaissances. Cette similarité est fondée sur la distance de Levenshtein, particulièrement adaptée pour prendre en compte les variantes de noms et les erreurs orthographiques. Dans nos expérimentations, nous avons retenu comme candidate une entité de la base de connaissances si sa forme ou l'une de ses variantes possède une distance avec la mention d'entité ≤ 2. Cette sélection s'appuie sur une structure de données de type BK-tree (Burkhard et Keller, 1973) pour une mise en œuvre efficace en termes de temps de calcul.

Les entités candidates sont également filtrées en fonction du type identifié de la mention d'entité (i.e. PERSONNE, LIEU ou ORGANISATION).

3.2 Sélection de la meilleure entité candidate

L'objectif de cette étape est d'identifier l'entité à laquelle la mention d'entité cible fait référence au sein de l'ensemble des entités candidates générées. Cette identification repose sur un classifieur entraîné à partir de mentions d'entités désambiguïsées.

Pour opérer cette classification, chaque entité candidate est décrite par un ensemble de traits :

– quatre traits binaires précisent la stratégie utilisée pour produire l'entité candidate ;

2. Dans ce qui suit, nous ne ferons référence qu'à la mention d'entité pour ne pas alourdir le propos mais ce vocable couvre également ses formes alternatives trouvées dans le texte source.

– deux scores caractérisent la similarité entre le contexte de la mention d'entité cible, au sens de son environnement textuel, et les éléments de contexte associés à l'entité candidate. Le premier score se focalise sur le contexte lexical. Dans le cas de la mention cible, il s'agit assez naturellement du texte entourant la mention (on considère ici la totalité du document contenant la mention comme contexte textuel). Le contexte lexical de l'entité candidate est quant à lui constitué par la page Wikipédia qui lui est associée, lorsqu'il en possède une. Plus précisément, les deux contextes lexicaux respectifs sont transformés en vecteurs, selon un modèle vectoriel standard s'appuyant sur une pondération *tf-idf*, et leur similarité est évaluée en appliquant la mesure Cosinus entre ces deux vecteurs. Le second score est axé quant à lui sur les entités apparaissant dans le voisinage de la mention cible et de l'entité candidate. Les entités environnantes de l'entité candidate sont obtenues en suivant les relations directes impliquant cette entité dans la base de connaissances de référence. Comme pour le contexte lexical, cet ensemble d'entités est transformé en un vecteur et la similarité de contexte relationnel de l'entité candidate avec la mention cible est évaluée en appliquant la mesure Cosinus entre ce vecteur et le vecteur du contexte textuel de la mention cible (on utilise le vecteur global du document au lieu de se restreindre à un vecteur construit seulement sur les entités pour contourner des erreurs possibles d'identification des entités nommées dans le texte).

Le classifieur statistique binaire s'appuyant sur ces traits est appliqué à des couples (mention, entité candidate) et permet donc d'évaluer si une entité candidate est effectivement le référent d'une mention d'entité. Les données disponibles pour l'entraînement de ce classifieur sont constituées uniquement de couples (mention, entité de référence), c'est-à-dire d'exemples positifs. Nous construisons les exemples négatifs en générant les entités candidates à partir de la mention cible et en produisant les couples (mention, entité candidate), avec entité candidate $\neq$ entité de référence. Les entités candidates générées pouvant former un ensemble assez large (entre 1 et 460 dans nos expériences), nous procédons à un sous-échantillonnage des exemples négatifs en limitant leur nombre à 10 fois le nombre d'exemples positifs. Chaque décision produite par le classifieur est accompagnée d'une probabilité. Au final, l'entité candidate de plus forte probabilité est sélectionnée comme référent de la mention d'entité considérée.

La tâche de désambiguïsation des entités est traditionnellement associée à la capacité de détecter les mentions d'entité ne faisant référence à aucune entité de la base de connaissances de référence (cas des entités dites NIL). Dans notre approche, l'absence d'entité de référence est décidée lorsqu'aucune entité candidate n'est générée ou lorsque toutes les entités candidates générées sont écartées par le classifieur ci-dessus.

De façon expérimentale, nous avons testé plusieurs modèles de classifieurs, en validation croisée sur le corpus d'entraînement : en raison de la forme particulière des vecteurs de traits (peu de traits, qui ont des comportements différents, certains étant binaires, d'autres réels), nous avons privilégié des classifieurs du type Adaboost, Arbres de décision ou Forêts Aléatoires, mais nous avons également testé des SVM, à noyau linéaire ou RBF. Les meilleurs résultats étaient obtenus avec les modèles

Adaboost et SVM linéaires. Les résultats présentés dans cet article sont ceux obtenus avec Adaboost.

3.3 Score discriminant de désambiguïsation

Nous présentons dans cette section le nouveau trait que nous proposons de prendre en compte pour lier une mention à une entité, trait prenant la forme d'un score appelé *Score Discriminant de Désambiguïsation* (DDS, pour *Discriminative Disambiguation Score*). Le DDS rend compte de la probabilité pour une mention d'entité d'être désambiguïsée par une entité candidate donnée. Plus spécifiquement, il s'identifie à la probabilité a posteriori $P(candidate_i|mention)$ qu'une entité candidate soit la bonne désambiguïsation d'une mention, évaluée à partir du résultat d'un classifieur (Platt, 1998).

La nouveauté de notre approche tient dans le fait que le DDS repose sur l'apprentissage d'un classifieur pour chaque entité de la base de connaissances de référence. Dans la phase de désambiguïsation, la probabilité évoquée ci-dessus est donc calculée à partir d'un classifieur spécifique à chaque entité candidate (dans la mesure où des données existent pour son entraînement). Cette approche a précédemment été jugée difficile à mettre en œuvre (Fan *et al.*, 2015) à cause de ses exigences en termes calculatoires. De fait, l'une des difficultés de l'approche que nous proposons réside dans la capacité à apprendre de tels classifieurs pour des millions d'entités avec une bonne performance en termes de discrimination. Pour rendre cette approche réalisable, à la fois au niveau de la phase d'apprentissage et de la phase de désambiguïsation, nous nous sommes restreints à des classifieurs linéaires, en l'occurrence de type régression logistique. Il est à noter, sans que nous le détaillions ici, que nous avons obtenu des résultats similaires avec des SVM linéaires.

Pour chaque entité considérée, il est donc nécessaire de sélectionner un ensemble d'exemples positifs, un ensemble d'exemples négatifs, d'extraire les traits nécessaires à la construction de la représentation de chaque exemple et d'apprendre un classifieur à partir de tous ces exemples. Comme dans le cas du calcul du score de similarité lexicale (cf. Section 3.2), chaque exemple se voit associer une représentation de type « sac de mots » des éléments de contexte textuel destinés à le caractériser. Cette représentation prend la forme d'un vecteur, avec une pondération *tf-idf*, situé dans le même espace que celui défini pour le score de similarité lexicale globale.

Concernant les exemples positifs, le contexte textuel de chaque entité de la base de connaissances prend en compte les éléments suivants :

– le résumé de la page Wikipédia associée à l'entité ;

– les paragraphes contenant explicitement l'entité dans la page Wikipédia qui lui est associée ;

– les paragraphes issus d'autres pages Wikipédia contenant un lien (wikilink) pointant vers l'entité.

Comme suggéré par (Fan *et al.*, 2015), une stratégie directe de type « un contre tous » serait irréalisable en termes calculatoires. Pour résoudre ce problème, nous proposons trois approches différentes pour sélectionner les exemples négatifs :

– **DDS-Rand** : dans l'approche aléatoire, les exemples négatifs sont sélectionnés de façon aléatoire parmi les exemples positifs de toutes les autres entités de la base de connaissances. Aucune contrainte n'est imposée quant au caractère ambigu ou non de ces entités en termes de forme par rapport à l'entité considérée ;

– **DDS-Ambig** : dans l'approche ambiguë, les exemples négatifs sont sélectionnés aléatoirement à partir des exemples positifs d'entités ambiguës. Ces entités ambiguës sont générées de la même façon que les entités candidates lors du processus de désambiguïsation (cf. Section 3.1) : pour chaque forme connue dans la base de connaissances (forme normalisée ou variation) de l'entité considérée, les entités ambiguës sont les entités partageant cette forme ou ayant une forme proche (inclusion ou distance d'édition au niveau caractère ≤ 2). Comme l'ensemble des entités ambiguës peut être grand, les exemples négatifs sont sélectionnés aléatoirement en son sein ;

– **DDS-Ambig-NN** : dans l'approche dite des plus proches voisins, l'entité cible est représentée par le centroïde des vecteurs *tf-idf* correspondant à ses exemples positifs. Ses exemples négatifs sont sélectionnés en considérant les entités ambiguës (au sens de DDS-Ambig) dont les vecteurs sont les plus proches de ce centroïde selon la mesure de similarité Cosinus. Ces exemples négatifs sont considérés comme les plus discriminants dans la mesure où ce sont les plus ambigus par rapport à l'entité et les plus proches de l'hyperplan séparateur entre les exemples positifs et négatifs.

4 Évaluation

4.1 Corpus d'évaluation

Pour évaluer notre approche, nous utilisons les corpus d'évaluation provenant des campagnes d'évaluation TAC-KBP, d'une part pour les années 2009-2013 et d'autre part pour l'année 2015. Pour cette dernière, notre évaluation correspond à la tâche diagnostique monolingue, pour l'anglais, dans laquelle les mentions des entités dans les textes fournis en requête sont données en entrée : ce cadre forme ainsi une évaluation cohérente avec les autres années. Nous présentons, dans le tableau 1, les caractéristiques principales de ces différents corpus d'évaluation. Dans la campagne 2015, le but était de désambiguïser la totalité des entités d'un petit nombre de documents alors que pour les campagnes précédentes, le focus portait sur un nombre donné d'entités, en produisant pour chacune un document en contexte : ceci explique que, pour ces campagnes, le nombre d'entités est approximativement le même que le nombre de documents.

Pour les sessions 2009-2013, la base de connaissances de référence est extraite des informations structurées de Wikipédia (*infoboxes*), de façon similaire à *DBPedia* (Namee *et al.*, 2009). La base construite contient 818 741 entités, qui sont toutes associées à des pages Wikipédia. Dans la campagne 2015, la base de connaissances

	Nb. docs.	Nb. entités
	KB = Dbpedia	
TAC 2009	3 688	3 904
TAC 2010	2 231	2 250
TAC 2011	2 231	2 250
TAC 2012	2 016	2 226
TAC 2013	1 820	2 190
	KB = Freebase	
TAC 2015 entraînement	168	12 175
TAC 2015 test	167	13 587

Tableau 1. *Description des corpus d'évaluation utilisés pour la désambiguïsation d'entités*

est construite à partir de Freebase (Ji *et al.*, 2015). L'image de référence de Freebase considérée dans la campagne recèle 43 millions d'entités, mais un premier filtre préalable, spécifié dans les instructions de la campagne, est appliqué pour supprimer un grand nombre d'entités dont les types ne sont pas pertinents par rapport aux entités de la campagne (comme les titres d'œuvres, livres, films ou morceaux de musique ou des entités médicales), ce qui réduit le nombre d'entités à 8 millions (10 fois plus que dans la base de connaissances utilisée en 2009-2013). Parmi ces entités, seules 3 712 852 (46%) ont une page associée dans Wikipédia et possèdent donc une information de contexte textuel permettant de construire les données pour apprendre le DDS.

4.2 Évaluation des entités candidates

Le tableau 2 présente des statistiques sur les requêtes et les entités candidates générées : on note R l'ensemble des requêtes formées par les mentions d'entités en contexte, R_{NIL} l'ensemble des requêtes qui n'ont pas d'entités associées dans la base de connaissances, C l'ensemble des candidats, C_{NIL} l'ensemble des requêtes pour lesquelles aucun candidat n'est proposé, C_{AVG} le nombre moyen de candidats par requêtes et Rappel(C) le rappel sur les candidats, défini par le pourcentage des requêtes (non NIL) pour lesquelles le candidat attendu est présent parmi la liste des candidats. Les chiffres montrent que ce rappel est plutôt bon pour les corpus 2009-2013, en utilisant des stratégies de production des candidats relativement simples donnant un nombre raisonnable de candidats par requête (150 en moyenne). Pour le corpus 2015, la base de connaissances contient 10 fois plus d'entités et le nombre de candidats produits est également beaucoup plus élevé. De plus, le rappel sur les candidats est plus faible (77%) : une analyse des entités manquantes a montré que les variations des noms d'entités présentes dans Freebase ne sont pas assez complètes et devraient être enrichies pour une meilleure couverture (par exemple, les liens entre les adjectifs

Dbpedia						
Corpus	$\|R\|$	$\|R_{NIL}\|$	$\|C\|$	$\|C_{NIL}\|$	C_{AVG}	Rappel(C)
2009	3 904	2 229	208 060	949	70,41	84,0%
2010	2 250	1 230	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%
Freebase						
2015 train	12 175	3 215	5 844 592	1 282	458,08	76,0%
2015 test	13 587	3 379	6 141 369	1 255	480,32	77,6%

Tableau 2. *Statistiques sur les nombres de requêtes (R) et les candidats produits (C) pour les corpus DBpedia and Freebase (TAC 2009-2013 and 2015)*

de nationalité et les pays ne sont pas présents, comme *French* $\rightarrow$ *France*, alors qu'ils doivent être trouvés dans la désambiguïsation d'entités).

4.3 Évaluation intrinsèque du score DDS

L'extraction du contexte textuel (cf. Section 3.3) à partir des pages Wikipédia est effectuée pour la totalité des 818 741 entités de *DBPedia* et pour les 3 712 852 entités de Freebase qui ont une page Wikipédia. Un espace vectoriel de 169 647 dimensions, construit à partir de la totalité des pages Wikipédia, est utilisé pour servir de support à la représentation vectorielle des paragraphes extraits, selon un modèle *tf-idf*. Pour *DBPedia*, un total de 32 939 218 exemples sont générés, alors que pour Freebase, on construit ainsi 97 157 120 exemples. Une entité est ainsi associée à 26,2 exemples en moyenne avec, en pratique, un nombre d'exemples compris entre 1 et 119 197. Les entités candidates pour TAC 2009-2013 (cf. tableau 2, colonne $|C|$) représentent 41 313 entités uniques, alors que pour TAC 2015, on a 124 456 entités uniques (en cumulant les entités candidates des corpus d'entraînement et de test). Pour chaque entité candidate de DBPedia ou de Freebase[3], nous entraînons un classifieur linéaire selon l'approche décrite à la Section 3.3 : dans nos expériences, nous utilisons un modèle de régression logistique avec régularisation L2 en nous appuyant sur l'implémentation de la bibliothèque Liblinear[4].

3. Notons que nous nous restreignons ici à l'apprentissage des classifieurs sur ces entités pour des raisons de simplicité pour l'évaluation mais que cela ne biaise pas l'évaluation. De façon opérationnelle, on pourrait entraîner un classifieur pour toutes les entités de la base de connaissances et avoir les mêmes résultats, puisque cet apprentissage n'est pas dépendant des requêtes.
4. `https://www.csie.ntu.edu.tw/~cjlin/liblinear`

	Dbpedia			Freebase		
	Min.	**Max.**	**Moy.**	**Min.**	**Max.**	**Moy.**
DDS-Rand	0,003	109,59	1,55	0,002	49,29	1,13
DDS-Ambig	0,01	398,47	11,65	0,006	270,45	6,49
DDS-Ambig-NN	0,027	2551,62	146,88	0,014	2102,31	85,49

Tableau 3. *Temps minimum, maximum et moyen (en secondes) nécessaire à chaque approche pour sélectionner les exemples négatifs et entraîner un classifieur pour une entité, pour DBpedia et Freebase*

	Précision	**Rappel**	**F-score**
DDS-Rand	$0,987 \pm 0,015$	$0,969 \pm 0,042$	$0,977 \pm 0,028$
DDS-Ambig	$0,963 \pm 0,050$	$0,919 \pm 0,112$	$0,937 \pm 0,086$
DDS-Ambig-NN	$0,954 \pm 0,058$	$0,798 \pm 0,188$	$0,857 \pm 0,151$

Tableau 4. *Résultats en validation croisée des classifieurs entraînés sur 26 819 entités ayant au moins 100 exemples positifs.*

Nous présentons dans le tableau 3 les temps minimum, maximum et moyen, en secondes, pour entraîner un tel classifieur selon les différentes approches. *DDS-Rand* est l'approche la plus simple en termes de traitements pour la sélection des exemples négatifs : elle est donc beaucoup plus rapide. *DDS-ambig-NN* est l'approche la plus longue car la sélection des exemples négatifs demande le calcul de la distance entre le centroïde des exemples de l'entité et chaque vecteur *tf-idf* des exemples de ses entités ambiguës.

Pour tester la pertinence des scores DDS, nous avons d'abord sélectionné un sous-ensemble de 26 819 entités de la base de connaissances extraite de DBPedia ayant au moins 100 exemples positifs et évalué la qualité des classifieurs entraînés pour ces entités en utilisant une validation croisée à 5 plis (*5-fold cross validation*). Les résultats de ces classifieurs, pour les mesures standard de précision, rappel et F-score, sont présentés dans le tableau 4 et montrent que ces classifieurs permettent effectivement de bien différencier une entité particulière d'autres entités. Les résultats sont un peu moins bons pour *DDS-ambig-NN* parce que, dans ce cas, nous sélectionnons spécifiquement les exemples négatifs les plus proches, ce qui rend la tâche de désambiguïsation plus compliquée : différencier une entité de la totalité des autres est globalement plus simple que la différencier des entités qui lui ressemblent le plus.

4.4 Résultats sur la désambiguïsation des entités

Dans cette section, nous présentons les résultats obtenus avec les scores DDS sur la tâche complète de désambiguïsation d'entités nommées. Nous comparons les résultats du système *baseline* présenté à la Section 3.2 avec les résultats obtenus en ajoutant chaque score DDS à l'ensemble des traits associés à chaque candidat, pour une entité requête donnée (le score DDS est donné par le classifieur associé à l'entité candidate).

Les modèles discriminants par entité sont calculés à partir de contextes textuels additionnels par rapport au modèle de base. Pour vérifier l'apport spécifique de ces modèles discriminants par rapport au simple ajout de plus de contextes textuels, nous considérons également un score égal à la similarité Cosinus entre le centroïde des exemples positifs pour une entité candidate et le contexte textuel de la mention d'entité à désambiguïser : ce score est nommé DDS-baseline.

Corpus d'entraînement et de test Pour TAC 2015, nous utilisons la séparation officielle entre les corpus d'entraînement et de test. Pour les campagnes TAC 2009-2013, aucun corpus d'entraînement explicite n'était fourni. Pour évaluer notre approche, nous utilisons dans ce cas, pour chaque année, les données des autres années comme corpus d'entraînement.

Mesures d'évaluation L'évaluation de la désambiguïsation est faite uniquement sur la première entité renvoyée par le système. Nous utilisons les mesures standard de précision/rappel/F-score sur trois critères : la bonne reconnaissance de l'entité de référence lorsqu'elle existe (*link*), la bonne reconnaissance d'une mention sans entité de référence associée (*nil*) et les résultats combinés des deux cas (*all*). Ces mesures correspondent aux mesures nommées *strong_link_match*, *strong_nil_match* et *strong_all_match* dans la campagne d'évaluation TAC 2015 (Ji *et al.*, 2015). Le type des entités nommées n'est pas pris en compte dans cette évaluation.

Résultats et analyse Nous présentons dans le tableau 5 les résultats en F-score pour les différentes approches, pour les mesures *strong_nil_match*, *strong_link_match*, *strong_all_match*. Ces résultats montrent que l'inclusion du score DDS dans l'ensemble des traits utilisés pour la sélection de la meilleure entité candidate améliore nettement les valeurs de F-score. Les meilleurs résultats sont obtenus avec le modèle *DDS-Ambig-NN* pour les corpus TAC 2009-2013, alors que le score *DDS-Ambig* donne les meilleurs résultats pour TAC 2015. On remarque que, même si le modèle *DDS-Ambig-NN* ne donne pas les meilleurs résultats pour la tâche de classification, son utilisation dans le système complet améliore les résultats globaux, ce qui tend à montrer que ce modèle apprend des informations discriminantes offrant une meilleure complémentarité avec l'information fournie par les autres traits. Par ailleurs, les résultats varient de façon significative selon les années, ce que l'on peut interpréter en première analyse comme un reflet des différences de difficulté propres aux données de ces campagnes : les scores des meilleurs participants à chacune de ces campagnes suivent en effet les mêmes variations.

		2009	2010	2011	2012	2013	2015
Baseline	*nil*	0,851	0,863	0,808	0,649	0,801	0,668
	link	0,707	0,743	0,645	0,441	0,717	0,588
	all	0,795	0,813	0,735	0,533	0,761	0,601
DDS-Baseline	*nil*	0,851	0,859	0,807	0,649	0,800	0,667
	link	0,709	0,736	0,639	0,446	0,705	0,603
	all	0,796	0,808	0,734	0,535	0,754	0,611
DDS-Rand	*nil*	0,856	0,858	0,817	**0,651**	0,801	0,679
	link	0,720	0,751	0,646	0,436	0,704	**0,659**
	all	0,803	0,813	0,741	0,531	0,753	0,654
DDS-Ambig	*nil*	0,858	0,867	0,812	0,643	0,799	**0,694**
	link	0,730	0,762	0,647	0,454	0,722	0,654
	all	0,808	0,824	0,741	0,537	0,763	**0,656**
DDS-Ambig-NN	*nil*	**0,874**	**0,884**	**0,821**	0,649	**0,82**	0,687
	link	**0,754**	**0,796**	**0,663**	**0,468**	**0,756**	0,644
	all	**0,828**	**0,848**	**0,752**	**0,547**	**0,789**	0,646

Tableau 5. *Résultats (F-score) obtenus sur TAC 2015 en ajoutant les scores DDS, selon les scores strong_nil_match (en haut), strong_link_match (au milieu) et strong_all_match (en bas). Les meilleurs scores sont en gras.*

Pour étudier plus avant l'influence du score DDS, nous avons testé une approche de désambiguïsation exploitant ce seul score : comme pour l'approche générale, un classifieur final est entraîné mais celui-ci utilise le score DDS comme seul trait (ce qui permet d'apprendre automatiquement un seuil sur la valeur de ce score pour la décision de classification). Le tableau 6 présente les résultats obtenus avec cette méthode, pour chaque variante de score DDS. À l'exception des corpus 2011 et 2015, le score *DDS-Ambig-NN* donne les meilleurs résultats et n'est pas très éloigné des scores obtenus par l'approche *baseline*. Pour le corpus 2012, le score DDS seul permet d'avoir un F-score *strong_all_match* de 63,8%, ce qui est même meilleur que le score obtenu en le combinant avec les autres traits.

Enfin, nous présentons une comparaison des résultats que nous avons obtenus sur ces corpus avec les résultats des autres participants aux campagnes, dans le tableau 7. Nous ne donnons les résultats que pour 2009, 2010 et 2015, seules années où la mesure officielle (*micro-average KB accuracy*) correspond à la mesure que nous utilisons dans nos expérimentations. Pour 2011-2013, la mesure est un score $B^3 + F1$, qui prend en compte un regroupement des mentions NIL faisant référence à la même entité (qui n'est pas dans la base de connaissances), alors que nous n'effectuons pas cette étape dans notre système. Notons aussi que le score officiel de TAC 2015 prend en compte le type de l'entité (même pour les entités NIL) alors que nous ne considérons que la désambiguïsation.

		2009	2010	2011	2012	2013	2015
DDS-Baseline	*nil*	0,749	0,718	0,655	0,639	0,634	0,405
	link	0,289	0,222	0,278	0,220	0,182	0,002
	all	0,609	0,560	0,493	0,489	0,470	0,245
DDS-Rand	*nil*	0,828	0,838	**0,771**	0,818	0,757	0,611
	link	0,622	0,687	**0,546**	0,156	0,585	**0,508**
	all	0,749	0,772	**0,670**	0,537	0,672	**0,541**
DDS-Ambig	*nil*	0,815	0,833	0,662	0,565	0,748	**0,665**
	link	0,568	0,666	0,243	0,221	0,560	0,443
	all	0,730	0,768	0,481	0,395	0,667	0,517
DDS-Ambig-NN	*nil*	**0,840**	**0,849**	0,756	**0,754**	**0,772**	0,633
	link	**0,625**	**0,703**	0,429	**0,446**	**0,614**	0,433
	all	**0,771**	**0,797**	0,645	**0,638**	**0,708**	0,503

Tableau 6. *Résultats (F-score) obtenus sur TAC 2015 en utilisant les scores DDS seuls, selon les scores strong_nil_match (en haut), strong_link_match (au milieu) et strong_all_match (en bas). Les meilleurs scores sont en gras.*

Pour les corpus 2009 et 2010, notre approche avec le score *DDS-Rand* serait classée seconde sur 18 et troisième sur 21. Pour 2015, notre système a un score correct qui le place au-dessus de la médiane, mais reste loin du meilleur système : comme nous l'avons remarqué, notre système *baseline* nécessiterait un enrichissement de la base de connaissances et une amélioration de la première étape de production des candidats qui permettrait d'avoir une meilleure base pour la seconde étape. Néanmoins, la tendance générale de nos résultats donne une base solide qui indique l'intérêt du score DDS que nous proposons pour la désambiguïsation des entités nommées.

	2009	2010	2015
Nb. of teams	18	21	10
Median	0,670	0,683	0,455
Min.	0,0085	0,345	0,030
Max.	0,822	0,864	0,737
Baseline	0,783 (8)	0,798 (9)	0,601 (5)
DDS-Baseline	0,791 (6)	0,810 (8)	0,611 (5)
DDS-Rand	**0,803 (2)**	**0,821 (3)**	0,654 (4)
DDS-Ambig	0,794 (5)	0,815 (6)	**0,656 (4)**
DDS-Ambig-NN	0,795 (3)	0,820 (4)	0,641 (4)

Tableau 7. *Comparaison de notre approche avec les résultats officiels des campagnes.*

5 Conclusion

Nous proposons dans cet article un nouveau critère pour la désambiguïsation des entités nommées, fondé sur une approche supervisée pour apprendre des modèles discriminants pour chaque entité d'une base de connaissances de référence. En intégrant ce nouveau critère dans un système général de désambiguïsation d'entités (qui combine également plusieurs autres traits communément utilisés), nous montrons que les résultats, exprimés en pourcentages, s'améliorent en moyenne de 4 points. Nous validons l'approche en évaluant notre système de désambiguïsation d'entités sur plusieurs corpus de référence provenant de différentes campagnes d'évaluation et utilisant des bases de connaissances différentes. Nous comparons plus précisément trois modèles pour la construction des modèles discriminants, intégrant des stratégies différentes pour la sélection des exemples négatifs : les modèles *DDS-Rand* et *DDS-Ambig* permettent d'améliorer le système de base sans trop accroître le temps de traitement, en gardant une complexité linéaire. L'approche *DDS-Ambig-NN* donne les meilleurs résultats mais au prix d'un temps de traitement plus élevé. Enfin, nous avons montré que l'apprentissage de classifieurs binaires individuels pour toutes les entités d'une base de connaissances de grande taille est possible et intéressant.

Deux pistes sont actuellement envisagées pour améliorer la qualité de notre système de désambiguïsation : d'une part, le système de production des candidats doit évoluer pour augmenter le rappel sur les candidats, en intégrant plus de connaissances extérieures dans cette étape ; d'autre part, au niveau des scores DDS, nous prévoyons d'utiliser des représentations vectorielles denses des mots (*word embeddings*) pour représenter les textes des exemples positifs et négatifs, ce type de représentation ayant prouvé son efficacité dans de nombreuses tâches de classification.

6 Bibliographie

Burkhard W. A., Keller R. M., "Some Approaches to Best-match File Searching", *Communications of the ACM*, vol. 16, n° 4, p. 230-236, April, 1973.

Cao Z., Tao Q., Tie-Yan L., Ming-Feng T., Hang L., "Learning to Rank: From Pairwise Approach to Listwise Approach", *24th International Conference on Machine Learning (ICML 2007)*, Corvalis, Oregon, USA, p. 129-136, 2007.

Cassidy T., Chen Z., Artiles J., Ji H., Deng H., Ratinov L.-A., Zheng J., Han J., Roth D., "CUNY-UIUC-SRI TAC-KBP2011 Entity Linking System Description", *Text Analysis Conference (TAC 2011)*, 2011.

Cucerzan S., "Large-Scale Named Entity Disambiguation Based on Wikipedia Data", *2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, Prague, Czech Republic, p. 708-716, 2007.

Dredze M., McNamee P., Rao D., Gerber A., Finin T., "Entity Disambiguation for Knowledge Base Population", *23rd International Conference on Computational Linguistics (COLING'10)*, Beijing, China, p. 277-285, 2010.

Fan M., Zhou Q., Zheng T. F., "Distant Supervision for Entity Linking", *29th Pacific Asia Conference on Language, Information and Computation (PACLIC 29)*, Shanghai, China, p. 79-86, 2015.

Han X., Zhao J., "NLPR_KBP in TAC 2009 KBP Track: A Two-Stage Method to Entity Linking", *Text Analysis Conference (TAC 2009)*, 2009.

Ji H., Nothman J., Hachey B., "Overview of TAC-KBP2014 Entity Discovery and Linking Tasks", *Text Analysis Conference (TAC 2014)*, 2014.

Ji H., Nothman J., Hachey B., Florian R., "Overview of TAC-KBP2015 Tri-lingual Entity Discovery and Linking", *Text Analysis Conference (TAC 2015)*, 2015.

Khalid M. A., Jijkoun V., de Rijke M., "The Impact of Named Entity Normalization on Information Retrieval for Question Answering", *in* C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, R. W. White (eds), *Advances in Information Retrieval: 30th European Conference on IR Research (ECIR 2008), Glasgow, UK*, Springer Berlin Heidelberg, p. 705-710, 2008.

Lehmann J., Monahan S., Nezda L., Jung A., Shi Y., "LCC Approaches to Knowledge Base Population at TAC 2010", *Text Analysis Conference*, 2010.

Ling X., Singh S., Weld D., "Design Challenges for Entity Linking", *Transactions of the Association for Computational Linguistics (TACL)*, vol. 3, p. 315-328, 2015.

Mihalcea R., Csomai A., "Wikify! Linking Documents to Encyclopedic Knowledge", *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, ACM, Lisbon, Portugal, p. 233-242, 2007.

Moro A., Raganato A., Navigli R., "Entity Linking meets Word Sense Disambiguation: a Unified Approach", *Transactions of the Association for Computational Linguistics (TACL)*, vol. 2, p. 231-244, 2014.

Namee P. M., Simpson H., Dang H. T., "Overview of the TAC 2009 Knowledge Base Population Track", *Text Analysis Conference (TAC 2009)*, 2009.

Platt J. C., "Fast Training of Support Vector Machines Using Sequential Minimal Optimization", *Advances in Kernel Methods - Support Vector Learning*, MIT Press, January, 1998.

Shen W., Jianyong W., Ping L., Min W., "LINDEN: Linking Named Entities with Knowledge Base via Semantic Knowledge", *21st International Conference on World Wide Web (WWW'12)*, Lyon, France, p. 449-458, 2012.

Shen W., Wang J., Han J., "Entity Linking With a Knowledge Base: Issues, Techniques, and Solutions", *Transactions on Knowledge and Data Engineering*, 2015.

Varma V., Bharath V., Kovelamudi S., Bysani P., GSK S., N K. K., Reddy K., Kumar K., Maganti N., "IIT Hyderabad at TAC 2009", *Text Analysis Conference (TAC 2009)*, 2009.

Zhang W., Chuan S. Y., Jian S., Lim T. C., "Entity Linking with Effective Acronym Expansion, Instance Selection and Topic Modeling", *Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI-11)*, Barcelona, Catalonia, Spain, p. 1909-1914, 2011.

Zhang W., Jian S., Lim T. C., Ting W. W., "Entity Linking Leveraging: Automatically Generated Annotation", *23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, p. 1290-1298, 2010.

Zheng Z., Xiance S., Fangtao L., Y C. E., Xiaoyan Z., "Entity Disambiguation with Freebase", *2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT'12)*, Macau, China, p. 82-89, 2012.