Apprentissage supervisé sur ressources encyclopédiques pour l'enrichissement d'un lexique de noms propres destiné à la reconnaissance des entités nommées

Nadia Okinina[1], Damien Nouvel[1,2], Nathalie Friburger[1], Jean-Yves Antoine[1]
(1) Université François Rabelais Tours, LI, 3 place Jean Jaurès, 41 000 Blois
(2) ALPAGE, INRIA Roquencourt
Prenom.Nom@univ-tours.fr

RÉSUMÉ

Cet article présente une méthode hybride d'enrichissement d'un lexique de noms propres à partir de la base encyclopédique en ligne Wikipedia. Une des particularités de cette recherche est de viser l'enrichissement d'une ressource existante (Prolexbase) très contrôlée décrivant finement les noms propres. A la différence d'autres travaux destinés à la reconnaissance des entités nommées, notre objectif est donc de réaliser un enrichissement automatique de qualité. Notre approche repose sur l'utilisation en pipe-line de règles déterministes basées sur certaines informations DBpedia et d'une catégorisation supervisée à base de classifieur SVM. Nos résultats montrent qu'il est ainsi possible d'enrichir un lexique de noms propres avec une très bonne précision.

ABSTRACT

Supervised learning on encyclopaedic resources for the extension of a lexicon of proper names dedicated to the recognition of named entities

This paper concerns the automatic extension of a lexicon of proper names by means of a hybrid mining of Wikipedia. The specificity of this research is to focus on the quality of the added lexical entries, since the mining process is supposed to extend a controlled existing resource (Prolexbase). Our approach consists in the successive application of deterministic rules based on some specific information of the DBpedia and of a supervised classification with a SVM classifier. Our experiments show that it is possible to extend automatically such a lexicon without adding a perceptible noise to the resource.

MOTS-CLÉS : reconnaissance des entités nommées, lexique de nom propre, enrichissement automatique de lexique, Wikipedia, règles, classification supervisée, machine à vecteurs de support, SVM.
KEYWORDS: named entities recognition, proper names lexicon, automatic extension of lexicon, Wikipedia, rules, supervised classification, support vector machines, SVM

1 Introduction

Les systèmes de recherche d'information langagière peuvent faire appel à diverses ressources afin de déterminer quels sont les éléments dignes d'intérêt. En particulier, la Reconnaissance des Entités Nommées (REN) nécessite des lexiques extensifs afin de détecter les noms propres, notamment ceux de personnes, de lieux et d'organisations. Les lexiques utilisés dans le domaine comprennent entre plusieurs centaines de milliers et plusieurs millions d'entrées. Se pose dès lors la question de leur constitution.

Parmi les approches non-supervisées, le bootstrapping permet d'enrichir itérativement un lexique (Riloff, Jones, 1999; Dredze et al., 2010) par utilisation d'un corpus et extraction de motifs. Mais la constitution de corpus suffisamment volumineux peut être laborieuse, pour un résultat incertain. Il est possible d'éviter cette étape par interrogation de moteurs de recherche ou la fouille de pages HTML (Downey et al., 2007; Nadeau, 2007) : le Web constitue alors une ressource de grand taille que l'on peut interroger de manière ciblée. Ces dernières années, de nombreux projets ont conduit à la mise en ligne de données plus ou moins structurées (Open Data) ouvertes et donc accessibles à tous ; ces données pouvant être utilisées pour constituer des ressources. L'encyclopédie Wikipedia constitue un ensemble de données d'un grand intérêt pour la constitution de ressources pour la REN. Les travaux de (Bunescu et Pasca, 2006; Charton et Torres-Moreno, 2009) sur l'extraction de connaissances dans Wikipedia l'ont déjà montré. De notre côté, nous cherchons à alimenter une base de données de noms propres, Prolexbase (Tran et Maurel, 2006). Prolexbase est un dictionnaire relationnel multilingue de noms propres et de leurs dérivés ; les noms propres contenus dans cette base sont classés selon une typologie très fine et leur entrée est contrôlée manuellement. Notre objectif est de créer une méthode semi-supervisée d'enrichissement de Prolexbase : il s'agit de proposer de nouveaux noms propres en grande quantité et avec une grande fiabilité (bruit limité) afin de limiter le travail de supervision manuelle. Cette méthode repose sur la constitution contrôlée de jeux de données, l'implémentation de règles sélectionnant les informations DBpedia utiles et le paramétrage automatique d'un classifieur SVM. En résultat, nous obtenons des listes ordonnées de noms propres candidats à l'ajout dans Prolexbase.

Dans cet article, nous présenterons tout d'abord la base de données à enrichir, Prolexbase, et la source de données Wikipédia. Ensuite, nous décrirons la mise au point des jeux de données et le paramétrage du classifieur utilisé. Enfin, nous nous présenterons une évaluation de notre approche et les résultats obtenus sur chaque type de noms propres de Prolexbase.

2 Les ressources utilisées

2.1 Prolexbase : une ressource à enrichir

Prolexbase[1] est à la fois une ontologie et une base de données multilingue de noms propres qui décrit des noms propres et leurs dérivés appartenant au langage courant. Cette base ne

[1] Créé au LI (Université François Rabelais Tours), www.cnrtl.fr/lexiques/prolex/

comprend pas de termes de spécialité (médical, juridique etc.). 55000 noms propres (125000 formes fléchies) y sont classés selon 4 types principaux et leurs sous-types :

- **Anthroponymes** – *Anthroponymes individuels* tels que les célébrités, patronymes, prénoms ou pseudo-anthroponymes *; anthroponymes collectifs* tels que dynasties, ethnonymes, associations, ensembles, entreprises, institutions, organisations
- **Toponymes** – Toponymes territoriaux liés aux sociétés humaines (villes, états, régions, entités supranationales…), édifices, géonymes, voies de communication…
- **Ergonymes** – Objets, produits, pensées, vaisseaux, œuvres
- **Pragmonymes** – catastrophes, manifestations, fêtes, événements historiques, phénomènes météorologiques

Les entités recensées peuvent être liées entre elles par des relations de synonymie, méronymie, accessibilité ou d'expansion classifiante. Si nous ne cherchons à l'alimenter qu'en nouvelles entrées lexicales, il faut cependant tenir compte de la granularité et de la finesse de cette représentation, qui induit une grande qualité et interdit des ajouts approximatifs. Nous souhaitons ainsi enrichir Prolexbase à travers les catégories très classiques d'entités nommées telles que les personnes (Célébrités), les organisations (Ensembles, Entreprises, Institutions), les lieux (Edifices) mais aussi enrichir les catégories moins présentes dans Prolexbase (Œuvres, Marques, Manifestations, Catastrophes, Astronymes, Fêtes).

2.2 Wikipedia : une source d'enrichissement

Depuis quelques années, Wikipédia se présente comme la référence grand-public des encyclopédies en ligne. Elle est continuellement mise à jour par des contributeurs bénévoles ce qui fait sa grande richesse. Sa structure permet une exploitation facile par des moyens informatiques.

De cette encyclopédie, nous utilisons les articles, composés des éléments suivants :

- *Titre* – En cas d'homonymie, le titre contient entre parenthèses une précision (catégorisation sémantique) sur chaque entrée correspondant au terme ;
- *Infobox* (facultatif) – L'infobox est une manière concise de donner des informations résumées d'un article. Chaque infobox est créée d'après un template et a un nom unique. Le nombre d'infobox associées à un article est variable et seul un tiers des articles Wikipedia contient une infobox.
- *Texte* – De taille variable, il peut être créé selon un template ou non. Le premier paragraphe contient les informations essentielles sur le sujet de l'article. Il contient également des liens vers d'autres articles de Wikipédia ;
- *Eléments additionnels facultatifs* - Liste de notes ou de références, bibliographie, pointeurs vers des articles connexes et liens externes ;
- *Liste de catégories* – Un article est lié à des catégories, qui sont organisées dans une taxonomie sous forme de hiérarchie qui change très régulièrement. En moyenne, un article appartient à 2.68 catégories (Farina, 2010)

Grâce à la diversité des domaines qu'elle décrit, et son accessibilité par moyens informatiques, Wikipédia est largement utilisée en TAL, y compris dans le but d'enrichissement d'ontologies qui nous intéresse ici.

3 Classification supervisée pour l'enrichissement de Prolexbase

Notre approche pour l'enrichissement de l'ontologie Prolexbase est proche de celle de (Charton & Torres-Moreno, 2009). Comme dans cette approche, nous fouillons Wikipédia en utilisant conjointement des classifieurs numériques et des règles. Ce genre d'approche a de bons résultats en termes de description. Nous mettons en œuvre un apprentissage binaire par type.

Pour chaque type Prolexbase, nous avons donc constitué un corpus d'apprentissage qui comprend 2 parties :

- les **exemples positifs** sont sélectionnés manuellement au sein de Wikipédia, ou récupérés grâce à des correspondances existantes et non ambigües entre Prolexbase et Wikipédia.
- les **exemples négatifs** peuvent également être sélectionnés manuellement au sein de Wikipédia (pour les types peu représentés dans Wikipédia), ou par tirage au sort. La sélection manuelle a été inévitable dans deux situations : pour les célébrités, car elles sont très abondantes dans Wikipédia ; pour certaines entrées liées à l'art (manifestations, édifices, œuvres), les types possibles sont difficiles à distinguer, dans ce cas nous avons décidé de prendre comme exemple négatifs des entrées de types proches (par exemple, une manifestation peut faire un bon exemple négatif pour un édifice).

De cette manière, nous avons constitué des corpus d'apprentissages dédiés à chaque type d'entités nommées à classer. Afin que la classification soit efficace sur l'encyclopédie entière, pour un type donné, la proportion d'exemples positifs dans le corpus d'apprentissage doit être corrélée à sa représentativité dans Wikipédia. Ainsi, pour les entrées de Wikipédia dont le type est faiblement représenté, un sur-échantillonnage des exemples négatifs peut être nécessaire.

3.1 Choix du classifieur

Le choix d'un classifieur a été fait sur la base de tests préalables effectués avec différentes techniques : classifieurs bayésiens naïf et multinomial[2], régression Logistique, diverses variantes de classifieurs SVM, k plus proches voisins (k-ppv). En première approche, nos tests ont montré la possibilité d'utiliser un classifieur bayésien multimonial. Ceci était attendu pour deux raisons :

- les textes comparés (exemples positifs et négatifs) sont généralement de longueurs différentes, comme souvent sur Wikipedia ;
- la répétition d'un mot dans un texte peut être significative (si le mot « entreprise » se répète dans les catégories et dans le premier paragraphe, il y a plus de chance que l'article concerné traite d'une entreprise).

En élargissant les expériences, nous avons obtenus les meilleurs résultats par utilisation de

[2] Contrairement au classifieur naïf, le classifieur multinomial prend en compte la longueur du document et le nombre d'occurrences d'un mot dans le document

classifieurs SVM. Nous en avons utilisé deux variantes de noyau : linéaire et RBF (Radial Basis Function) (Denil, Trappenberg, 2010). En effet, il semble qu'un classifieur linéaire soit plus performant pour des types homogènes (selon les informations qui les caractérisent), tandis qu'il est nécessaire d'utiliser le modèle RBF lorsqu'il y a disparité dans les attributs qui caractérisent un type donné. Par exemple, un SVM linéaire se révèle insuffisant pour le type *ensemble*, qui comprend d'un côté les groupes de musique et de l'autre côté des équipes sportives (ces deux pôles se caractérisant par des attributs différents). Dans ce cas, les meilleurs résultats sont obtenus par utilisation d'un SVM avec noyau RBF. Ceci a aussi été observé dans d'autres travaux, qui soulignent la supériorité de ces noyaux sur des tâches de classification de textes (Ko et Seo, 2011).

Nous constatons par ailleurs une forte dégradation des performances lorsque les corpus sont bruités. La sélection des éléments pour constituer le corpus d'apprentissage est déterminante et doit être réalisée avec soin. En particulier, il est impossible d'utiliser Prolexbase sans filtrage pour constituer les exemples positifs, notamment par la présence d'homonymie. Voilà pourquoi nous ne sélectionnons que les exemples positifs qui ne sont pas ambigus.

3.2 Algorithme d'enrichissement

Notre approche combine règles d'extraction et apprentissage supervisé, étant donné qu'il est possible d'atteindre une bonne précision à l'aide de règles déterministes dans des cas bien identifiés. Plus précisément, nous avons adopté l'algorithme séquentiel suivant :

1. **Règles sur les infoboxes** - Tout d'abord, nous cherchons à détecter l'appartenance d'un article à un type Prolexbase à l'aide de règles sur les infoboxes, qui sont explicitement conçues pour catégoriser les articles. Nos règles présentent une très bonne précision pour tous les types, mais moins d'un tiers des articles Wikipédia contiennent une infobox. Pour les ensembles, on recherche des infoboxes contenant les expressions suivantes : 'equipe de, club de, club sportif, musique, (artiste), charte, groupe'.

2. **SVM** - Si l'article n'a pas été traité à l'aide des règles précédentes, nous passons à une classification par SVM. Cette méthode permet également de récupérer un grand nombre d'entités. Ses performances varient beaucoup suivant le type. Nous procédons donc à une adaptation des classifieurs propre à chaque type suivant les paramètres ci-dessous :
   - *SVM linéaire ou RBF* – Le SVM linéaire donne de meilleurs résultats pour les types célébrités et entreprises. Pour les autres types, nous utilisons un noyau RBF.
   - *Paramètre gamma* de contrôle de la forme de l'hyperplan séparateur des classes (RBF kernel) – Augmenter gamma, accroît le nombre de vecteurs de support. Nous gagnons en fidélité au corpus, mais perdons en capacité de généralisation.
   - *Prise en compte ou non des titres* – Certains types Prolexbase ont des titres très significatifs (institutions, catastrophes, manifestations, astronymes, voies...), d'autres non (célébrités, œuvres, entreprises, produits ...). Nos tests ont montré à l'opposé que les catégories apportaient toujours une information utile à la classification. Elles sont donc considérées par tous les classifieurs.
   - *Prise en compte ou non du premier paragraphe des articles* – Pour certains types Prolexbase, les catégories seules donnent assez d'informations pour classer l'article. C'est le cas des célébrités où l'ajout du premier paragraphe diminue les performances. Pour d'autres types, considérer le premier paragraphe améliore au

contraire les performances du SVM. Précisons que les titres, catégories et premiers paragraphes sont regroupés dans un unique attribut 'text' (vecteur de mots).

3. **Règles sur les titres** – Si l'article n'a pas été retenu par le SVM du fait d'un score insuffisant, nous appliquons des règles sur les titres. Ces règles sont moins robustes que celles sur les infoboxes à cause de l'homonymie entre les types (voir l'exemple du roman *Notre Dame de Paris*, par exemple). Il est donc important que ces règles arrivent en fin de chaîne de traitement. Par exemple, le SVM classifie les séries télévisées parmi les œuvres, alors que ce sont des produits dans la classification Prolexbase. L'erreur du classifieur est liée au fait que les films soient classés comme œuvres dans Prolexbase. Nous procédons donc à un post-traitement en regardant les mots-clefs comme *« série télévisée »* parfois précisées entre parenthèses dans le titre.

4. **Règles sur les catégories** – Enfin, si l'article n'a toujours pas été classé, nous appliquons des règles sur les catégories. Nos tests ont montré que celles-ci étaient peu robustes, et les catégories ont déjà été considérées par le SVM pour certains types. Pour ces raisons, ce filtre est employé en dernier recours.

Les entités qui ne sont retenues par aucune de ces étapes ne sont pas proposées pour l'ajout au lexique. Notons que si l'ordre des étapes décrites ci-dessus est imposé par l'algorithme de détection, chaque étape est facultative pour un type donné. Nous avons en effet conduit des expérimentations détaillées qui nous ont permis de déterminer quelles étapes étaient pertinentes pour chaque type considéré. Le paragraphe ci-dessous présente la synthèse de ces expérimentations.

4 Résultats et évaluation

Les tables 1 et 2 ci-dessous donnent la liste des entités récupérées par notre algorithme par type Prolexbase. Nous précisons à chaque fois le nombre d'entités, les méthodes utilisées pour chaque type, et la précision des différentes techniques mobilisées. Nous évaluons les résultats par adaptation d'une précision P@100. Dans le cas du SVM, les entrées sont ordonnées selon leur score, et nous évaluons les 100 premières. Pour les autres méthodes, 100 entrées sont tirées aléatoirement. Pour chaque type présenté par cette table, la précision est suffisamment haute pour que ces listes d'entités soient directement utilisables pour la reconnaissance d'entités nommées (au-dessus de 95%). Nous donnons à titre informatif également la précision obtenue avec des méthodes qui n'ont pas été retenues car trop peu performantes sur le type considéré. Les types absents de la table n'ont pas permis d'obtenir de bons résultats. La table montre que la méthode la plus facile à utiliser et la plus robuste est celle de règles sur les infoboxes, ceci grâce à la précision des infoboxes Wikipédia.

L'obtention de résultats peu bruités avec le classifieur SVM est plus difficile à atteindre. Ainsi, seuls 6 types sur les 11 présentés dans la table 1 ont obtenus des résultats à l'aide du classifieur SVM. La qualité des performances du classifieur SVM dépend de plusieurs facteurs :

- *Le nombre d'entités recherchées présentes dans Wikipédia.* L'augmentation de la base d'apprentissage améliore comme attendu les performances du classifieur. A part dans le cas très précis des catastrophes, une base d'apprentissage de plusieurs milliers d'exemples est nécessaire à l'obtention de bonnes performances.
- *La spécificité du vocabulaire employé dans les articles et les catégories de Wikipédia*

selon les types. Les célébrités et les entreprises ont été facilement détectées par le SVM linéaire, parce que les articles et les catégories de Wikipédia qui leur correspondent contiennent toujours les mêmes mots-clefs. Les autres types ont demandé l'utilisation du SVM à noyau RBF, dont le paramétrage est plus délicat. Ces types ne sont pas homogènes et contiennent des sous-ensembles dont chacun possède son propre champ lexical. Par exemple, les œuvres se divisent en œuvres littéraires, cinématographiques, picturales etc. ; chacun de ces sous-ensembles est caractérisé par l'emploi d'une terminologie différente.

Type Prolexbase	**Nombre d'entités récupérées à l'aide de règles sur les infoboxes**	**Nombre d'entités récupérées à l'aide de SVM**	**Nombre d'entités récupérées à l'aide de règles sur les titres ou bien les catégories**
Célébrités	112 632	100 472	0
Œuvres	44 958	2 958	5 947
Ensembles	11 148	0	11 019
Marques	16331	0	255
Entreprises	9 623	7 748	7 230
Institutions	1 757	265	4 373
Edifices	3 282	0	0
Manifestations	16 737	2 446	370
Catastrophes	525	283	0
Astronymes	461	0	144
Fêtes	126	0	0

Table 1 : listes utilisables dans la reconnaissance d'entités nommées

Type Prolexbase	**Nombre d'entités récupérés**	**Méthodes utilisées**	**Précision Infobox**	**Précision SVM**	**Précision catégorie ou titre**
Célébrités	213 004	Infobox, SVM linéaire	99 %	100 %	Catégories : 84 %
Œuvres	53 864	Infobox, SVM RBF, titre	100 %	99%	Titres : 100 %
Ensembles	22 157	Infobox, catégories	97 %	(50%)	Catégories : 99 %
Entreprises	24 601	Infobox, SVM linéaire, catégories	98 %	100 %	Catégories : 96 %
Marques	16 586	Infobox, titre	100 %		Titres : 100 %
Institutions	6 395	Infobox, SVM RBF, titre	100 %	100 %	Titres : 98 %
Edifices	3 282	Infobox	100%	(65%)	(Titres : 92%)
Manifestations	19 183	Infobox, SVM RBF	96 %	95%	
Catastrophes	808	Infobox, SVM RBF	100 %	100%	
Astronymes	605	Infobox, titres	100%	(83%)	Titres : 100%
Fêtes	126	Infobox	100%	(83%)	(Titres : 70%)

Table 2 : nombre d'entités trouvées par différents systèmes

La table 2 montre le nombre de résultats utilisables par chacun des filtres développés. Rappelons que si la précision est inférieure à 95%, nous décidons de ne pas retenir la méthode, d'où la valeur nulle du nombre d'entité récupérées. Cette table nous indique qu'il y a de grandes différences dans la manière de traiter les divers types Prolexbase. Nous voyons

également que, en nombre d'entités récupérées, les règles sur les infoboxes sont toujours plus productives que le SVM. Dans le cas des célébrités et des entreprises, les deux nombres sont proches : dans ce cas, le SVM apporte beaucoup, puisqu'il permet presque de doubler le nombre d'entités récupérées. Dans le cas des œuvres et des manifestations, la proportion d'entités récupérées à l'aide de SVM est très faible, mais elle permet tout de même de compléter l'extraction d'entrées candidates, ce qui n'est pas à négliger. Enfin, les catastrophes sont peu nombreuses par rapport aux autres types recherchés et le SVM en découvre tout de même 35%.

5 Conclusion

Cet article présente une méthode hybride combinant règles déterministes et apprentissage supervisé par machines à vecteurs de support pour l'enrichissement d'un lexique typé de noms propres à partir de Wikipédia. Les résultats montrent qu'il est possible d'atteindre un enrichissement important de Prolexbase avec un bruit limité. Notre objectif est désormais d'étudier l'influence de ces mises à jour du lexique sur les performances de nos systèmes de reconnaissance d'entités nommées.

Références

BUNESCU, R. C. ET PASCA, M. (2006) Using encyclopedic knowledge for named entity disambiguation, *Conference of the European Chapter of the Association for Computational Linguistics*.

CHARTON E., TORRES-MORENO J. (2009) Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées, Actes *TALN 2009*.

DENIL M., TRAPPENBERG T. (2010) Overlap versus Imbalance, Proc. *Canadian AI 2010*.

DOWNEY, D., BROADHEAD, M. ET ETZIONI, O. (2007) Overlap versus Imbalance , Proc.*Canadian AI 2010*. Locating complex named entities in web text, *International Joint Conference on Artifical Intelligence*, pages 2733–2739.

DREDZE, M., MCNAMEE, P., RAO, D., GERBER, A. ET FININ, T. (2010) Entity disambiguation for knowledge base population, *International Conference on Computational Linguistics*, 277–285, Beijing, Chine.

FARINA J. (2010) Assegnamento Automatico Di Macrocategorie Agli Articoli Di Wikipedia, *Tesi di Laurea Triennale*.

KO Y., SEO J., (2011) Issues and Empirical Results for Improving Text Classification , *Journal of Computing Science and Engineering*.

NADEAU, D. (2007) Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision, *thèse de doctorat*, University of Ottawa, Canada.

RILOFF, E., JONES, R. (1999) Learning dictionaries for information extraction by multi-level bootstrapping, *National Conference on Artificial Intelligence*, 474-479.TRAN M., MAUREL D. (2006) Prolexbase : Un dictionnaire relationnel multilingue de noms propres, *Traitement Automatique des Langues*, Vol. 47(3), 115-139.