Reconnaissance d'entités nommées : adaptation au domaine de la littérature française du XIXe-XXe siècle

Carmen Brando, **Francesca Frontini**, Alaa Abi Haidar and Jean-Gabriel Ganascia

8es Journées Internationales de Linguistique de Corpus - 2-4 sept 2015 Orleans (France)

UPMC PARISUNIVERSITAS

LIP6

Observatoire de la vie littéraire

Plan de la présentation

- Contexte du travail
- Etat de l'art sur les ENs et application au domaine de la littérature
- Corpus
- Expérience 1 et analyse
- Expérience 2 et analyse
- Conclusions
- Perspectives de recherche

LABEX OBVIL

L'**Observatoire de la vie littéraire (OBVIL)** entend développer toutes les ressources offertes par les applications informatiques pour examiner aussi bien la littérature française du passé que la plus contemporaine.

- Éditions numériques de textes littéraires et de critique
- Outils pour l'analyse du texte

http://obvil.paris-sorbonne.fr/

REN dans les textes litéraires

- La reconnaissance d’entités nommées (REN) est un enjeu fondamental pour la recherche en humanités numériques (HN).

- En littérature française, il est particulièrement important de repérer des entités telles que les auteurs, les personnages fictifs, les lieux géographiques et imaginaires, les titres d’ouvrages, les marqueurs temporels, entre autres.

- Actuellement, il existe peu de corpus français du domain litéraire annotés et disponibles en ligne.

Etat de l’art REN

Phases REN:

- ➔ identification
- ➔ classification (types d'entités: personne, lieu, organisation)

Approches:

- ➔ **supervisées** (état de l'art, haute précision, mais nécessitent un corpus d'entraînement)
- ➔ **non ou semi-supervisées** (mieux adaptables à différents domaines, utilisent souvent des dictionnaires)

REN dans des textes de critique

Corpus Critique du Labex OBVIL:

une large édition numérique de critique littéraire française.
http://obvil.paris-sorbonne.fr/corpus/critique/

Nécessité de repérer les citations de personnes (**auteurs**), des lieux (**centres culturels**), des organisations (**universités**)

Le coût élevé de l'annotation manuelle motive l'utilisation de méthodes automatiques pour l'exploration du corpus.

Expériences

L’adaptation à un nouveau domaine semble affecter négativement la performance des approches REN [Nadeau & Sekine 2007].

Comparer les performances de deux approches, **supervisée** et **non-supervisée**.

Évaluer de façon préliminaire l’adaptabilité de ces approches à un nouveau domaine tel que la littérature française du XIXe-XXe siècle.

L’approche UNERD

UNERD [Mosallam et al 2014] est un algorithme non supervisé (NSP) qui

- se sert d’un dictionnaire pour la détection de mentions certaines
- puis utilise leur contexte afin de classifier les mentions inconnues et ambiguës.

Le dictionnaire été adapté au domaine à partir du catalogue d’auteurs de la Bibliothèque Nationale de France (BnF) et de Dbpedia:

- ❏ **354.054 toponymes** y compris leurs noms alternatifs (noms anciens, noms officiels et noms vernaculaires)
- ❏ **3.941.206 noms d’auteurs** y compris leurs noms alternatifs (ex : M. Lamartine, Monsieur Lamartine, etc.)
- ❏ **14.068 organisations** (seuls les établissements éducatifs).

Apache OPENNLP

Le module REN de la suite Apache OPENNLP se base sur un algorithme supervisé (SP) de machine learning, et nécessite un corpus d'entraînement.

Dans ces expériences, on utilise le module avec les paramètres qui donnent en général les meilleurs résultats :

- algorithme d'apprentissage "perceptron"
- features ciblés pour le français (morphologie)

Corpus annotés manuellement

En collaboration avec les experts du Labex OBVIL, deux corpus ont été annotés manuellement pour l'entraînement et l'évaluation.

- "**Une Thèse sur le symbolisme**", Albert Thibaudet,1936 [THIBAUDET]
- "**L'Évolution créatrice**", Henri Bergson, 1907 [BERGSON]

Textes hétérogènes, au style caractérisé.

Bas degré de standardisation et de prédictibilité.
Richesse considérable en matière d'entités nommées.

Corpus annotés manuellement

Conventions:

➔ Les personnes sont annotées de manière à ce que leurs titres honorifiques soient inclus.

➔ Concernant la classification, nous nous sommes limités dans un premier temps au choix des classes dites Enamex : Personne (PER), Lieu (LOC), Organisation (ORG).

Première expérience - THIBAUDET

Évaluer la performance de chaque approche au sein du même corpus

NSP: annotation du THIBAUDET avec le dictionnaire adapté
SP: entraînement sur 3/4 du THIBAUDET et évaluation sur 1/4 (répété)

	Personne			Lieu			Organisation		
#Mentions	3410			382			30		
Mesure	P	R	F	P	R	F	P	R	F
NSP	0,78	**0,88**	0,83	0,46	**0,56**	0,50	0,01	0,03	0,14
SP	**0,86**	**0,88**	**0,87**	**0,66**	0,49	**0,57**	**0,13**	**0,18**	**0,15**

Première expérience: résultats

la méthode SP fournit les meilleurs résultats en ce qui concerne la précision.

Les rappels obtenus par les deux approches sont assez comparables, le rappel de la méthode NSP est particulièrement satisfaisant pour les personnes (0.88) et les lieux (0.56).

Le déséquilibre entre le nombre de mentions d'organisations (30) et les nombres de mentions de personnes (3410) et de lieux (382) mène à une performance décevante mais attendue vis-à-vis des organisations.

Analyse d'erreurs en détection

La méthode **NSP** choisit un nombre considérable de faux positifs lors de la détection.

- Premièrement, la manière dont UNERD identifie les entités candidates se base sur des critères uniquement orthographiques, tels que la présence de majuscules.
- Deuxièmement, une quantité considérable d'entités du corpus d'étude ne sont ni des personnes, ni des lieux, ni des organisations (ex : «la Magie de l'amour»).

Les erreurs les plus fréquentes faites par la méthode **SP** comprennent :

- l'oubli de mentions ayant un **contexte pauvre**
- l'identification de faux positifs à cause de la présence de **majuscules** au début de la phrase
- un **mauvais découpage** de la mention
  <person> Vigny </person> lisez dans **<person> Victor Hugo Melancholia </person> des Contemplations** .

Analyse d'erreurs en classification

Méthode NSP:

la classe **Personne contient un grand nombre de mentions correctement attribuées** grâce à l'utilisation du dictionnaire d'auteurs de la BnF.

Néanmoins, il y a des exceptions qui mènent à des problèmes de désambiguïsation. En particulier, des **entités qui sont inconnues du dictionnaire de personnes**, par exemple le personnage mythologique «Europe», mais qui existent dans le dictionnaire de lieux.

Méthode SP:

Les erreurs de classification faites par la méthode SP sont moindres, mais il est parfois possible de retrouver des **lieux classifiés comme des personnes** à cause d'un contexte ambigu.

Première expérience: conclusions

En général, un problème récurrent dans un texte de critique est la présence d'une quantité considérable de mentions de personnes qui apparaissent dans des contextes linguistiques difficilement distinguables d'autres classes.

[...] mais en marionnettes — les marionnettes de **<person> Ponson </person>** — conduites par des excitations extérieures . [...]

Deuxième expérience

Evaluer l'adaptation au domaine (critique littéraire > philosophie, science)
NSP: annotation du THIBAUDET avec le dictionnaire adapté
SP: entraînement BERGSON > test sur THIBAUDET

	Personne			Lieu			Organisation		
#Mentions	382			58			12		
Mesure	P	R	F	P	R	F	P	R	F
NSP	0,48	**0,87**	0,62	0,23	**0,83**	0,36	0	0	0
SP	**0,53**	0,71	0,61	**0,52**	0,71	**0,60**	**1**	**0,08**	**0,15**

Deuxième expérience: résultats

En général, **la précision pour les deux méthodes est faible** par rapport à la première expérience ; néanmoins, **la méthode SP reste performante**.

Les erreurs de mauvais étiquetage produites par les deux méthodes sont semblables à celles rencontrées dans le Thibaudet.

BERGSON contient des classes telles que les titres d'ouvrages, les êtres vivants (ex : bactéries, animaux) et les individus abstraits (ex : Dieu), créant des problèmes de faux positifs.

La méthode NSP surpasse de manière remarquable la méthode SP au niveau du rappel.

La méthode **SP, étant plus affectée par le contexte de la mention**, omet donc un nombre considérable de mentions.

rappel uniquement pour la détection NSP 0,82 - SP 0,71

Conclusion

En l’absence de corpus de référence (*gold standard*), l’objectif de notre travail est de faciliter l’annotation manuelle des nouveaux textes de critique française du XIXème et du XXème siècles.

Il semblerait en effet plus commode pour quelqu’un de corriger une annotation erronée que d’annoter le texte sans aucune aide.

De manière générale, il semble donc important de favoriser le rappel au détriment de la précision.

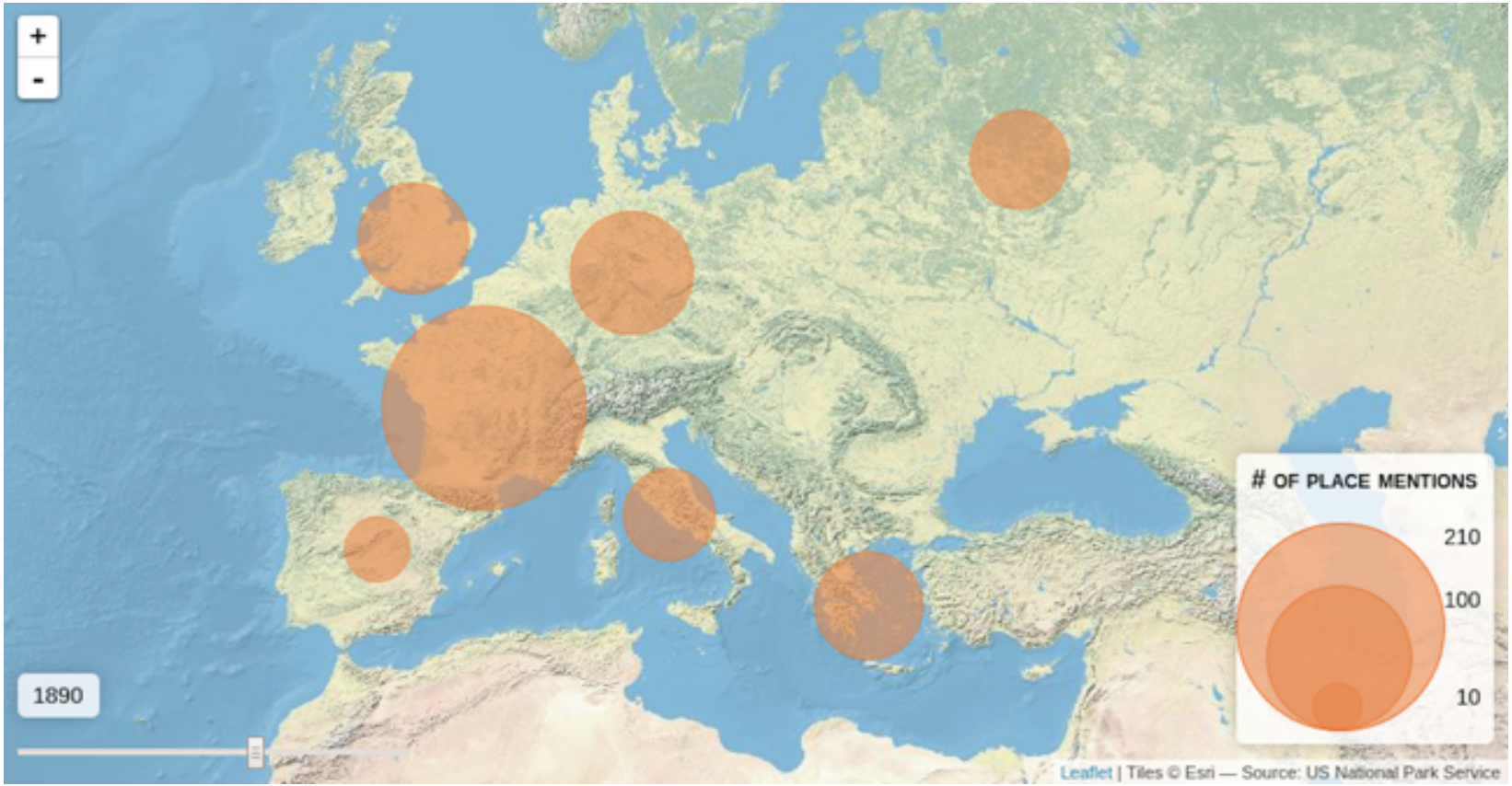
À long terme, l'idée est de produire un corpus de référence pour les ENs à partir du *Corpus Critique*.

Travaux en cours et perspectives

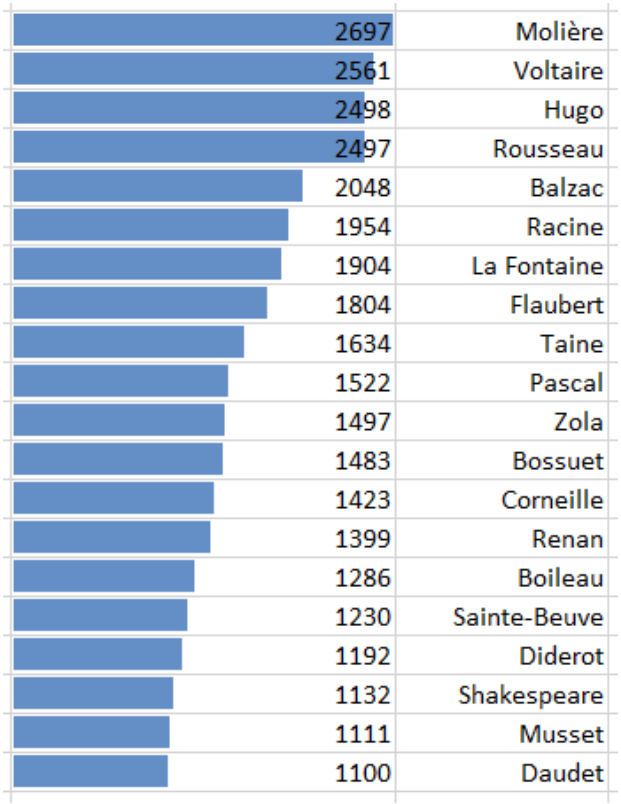
- REDEN [Frontini et al 2015, Brando et al 2015] outil pour le référencement des ENs (vers DBpedia, BnF)

- Nettoyage et analyse d'ensemble d'une partie du *Corpus Critique* pour découvrir les tendances littéraires et influences culturelles.

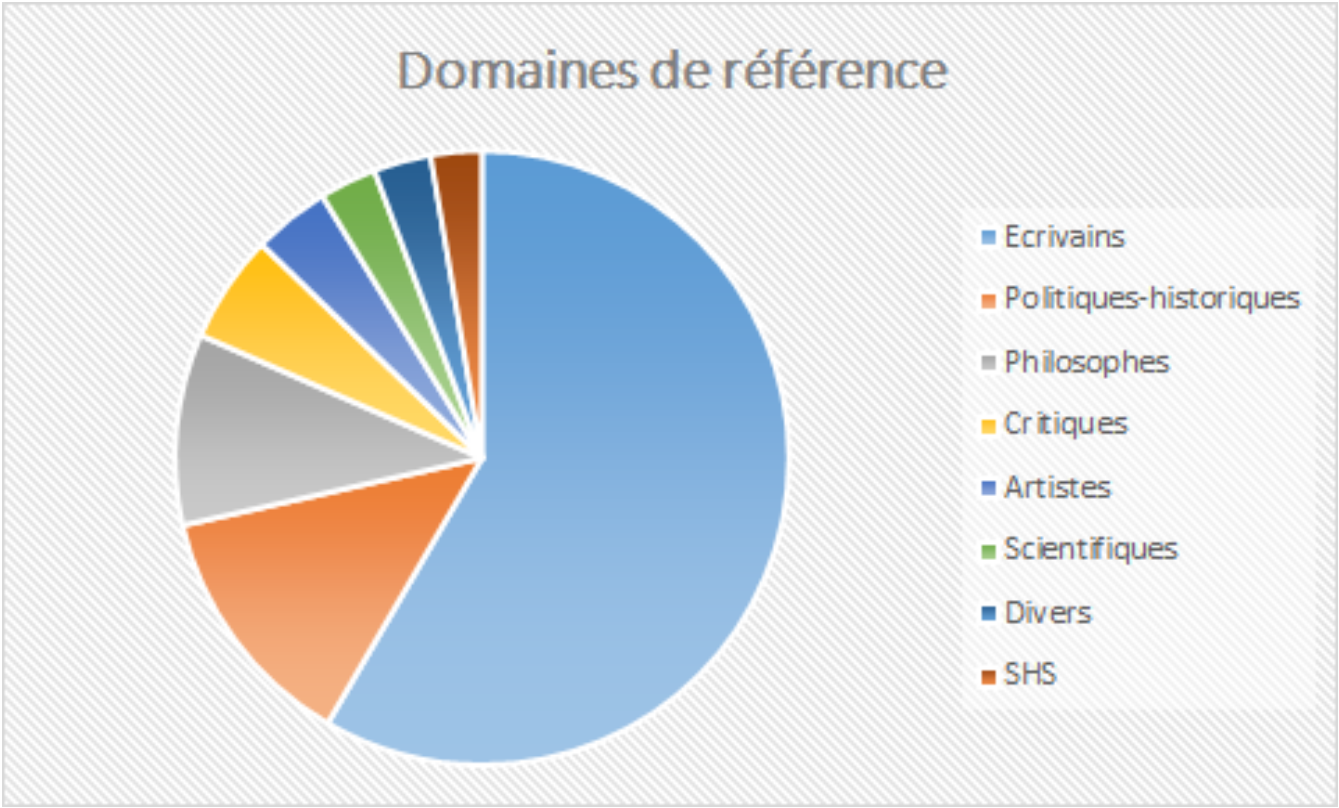
  Travaux par Marine Riguet - Labex OBVIL



Mentions de lieux par pays dans Corpus Critique dans les annés 1890.



Auteurs les plus cités dans le Corpus Critique.



Typologie des personnes citées dans le Corpus Critique.

Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., & Vaithyanathan, S. (2010, October). Domain adaptation of rule-based annotators for named-entity recognition tasks. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 1002-1012). Association for Computational Linguistics.

Brando, C., Frontini, F., Ganascia, J.G. (2015): Disambiguation of named entities in cultural heritage texts using linked data sets (accepted). In: Proceedings of the First International Workshop on Semantic Web for Cultural Heritage in Conjunction with 19th East-European Conference on Advances in Databases and Information Systems

Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press.

Frontini F, Brando C, Ganascia J-G, (2015) Semantic Web based Named Entity Linking for digital humanities and heritage texts, in Proceedings of the First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference, Portorož, Slovenia, June 1st, 2015, pp. 77-88, URL: http://ceur-ws.org/Vol-1364/paper9.pdf

Frontini, F., Brando, C., Ganascia, J.G. (2015): Domain-adapted named-entity linker using linked data. In: Proceedings of the 1st Workshop on Natural Language Applications: completing the puzzle in conjunction with the 20th International Conference on Applications of Natural Language to Information Systems, Passau, Germany, June 17-19, http://ceur-ws.org/Vol-1386/named_entity.pdf

Mosallam, Y., Abi-Haidar, A., & Ganascia, J. G. (2014). Unsupervised Named Entity Recognition and Disambiguation: An Application to Old French Journals. In Advances in Data Mining. Applications and Theoretical Aspects (pp. 12-23). Springer International Publishing.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26.

Poibeau, T., & Kosseim, L. (2001). Proper name extraction from non-journalistic texts. Language and computers, 37(1), 144-157.