Utilisation des méthodes de classification pour la construction automatique d'un corpus de Comptes Rendus de lecture

Chahinez Benkoussas*†, Patrice Bellot*†, Hussam Hamdan*†‡, Elodie Faath†

*LSIS, Aix-Marseille University

†CNRS / CLEO-OpenEdition, Aix-Marseille University

‡LIF, Aix-Marseille University

01 Juillet 2014

Problématique

- Recommandation automatique des livres d'OpenEdition dans les Sciences Humaines et Sociales.
- Comabinaison de l'analyse de contenu et le filtrage collaboratif.
- Considération du comportement des utilisateurs (navigation, lecture, ...) et leurs préférences (explicites, implicites). Comment?
- Analyse des Comptes Rendus de lecture (CR) –> les identifier automatiquement ...

Éléments clés

- Collecte automatique des Comptes Rendus de lecture (longs commentaires sur des livres).
- Utilisation des méthodes de classification thématique pour une classification en genre.
- Construire un corpus de critiques à partir des plateformes d'OpenEdition et du Web.

Éléments clés

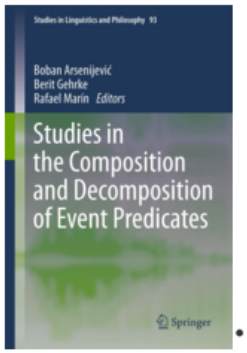
- Extrait d'un faux positif Compte Rendu de lecture.

–> Résumé de livre: Absence d'opinion.

Nouvelle publication : B. Arsenijević ; B. Gehrke ; R. Marín, (eds.)

Florence THILL

Nouvelle publication en linguistique : *Studies in the Composition and Decomposition of Event Predicates*



by **Boban Arsenijević**; **Berit Gehrke**; **Rafael Marín**, (eds.), Springer Verlag (Studies in Linguistics and Philosophy, 93), 2013, V, 79 p., 9 illus.

ISBN 978-94-007-5982-4

This book :

- Includes discussions of the processing aspects of the semantic ingredients of event predicates
- The role of scalar structures is shown to be more complex than traditionally assumed (one scale, simple mappings)
- Places the semantic entailments of event predicates and their more fine-grained components as the center of its study

This detailed, perceptive addition to the linguistics literature analyzes the semantic components of event predicates, exploring their fine-grained elements as well as their agency in linguistic processing. The papers go beyond pure semantics to consider their varying influences of event predicates on argument structure, aspect, scalarity, and event structure.

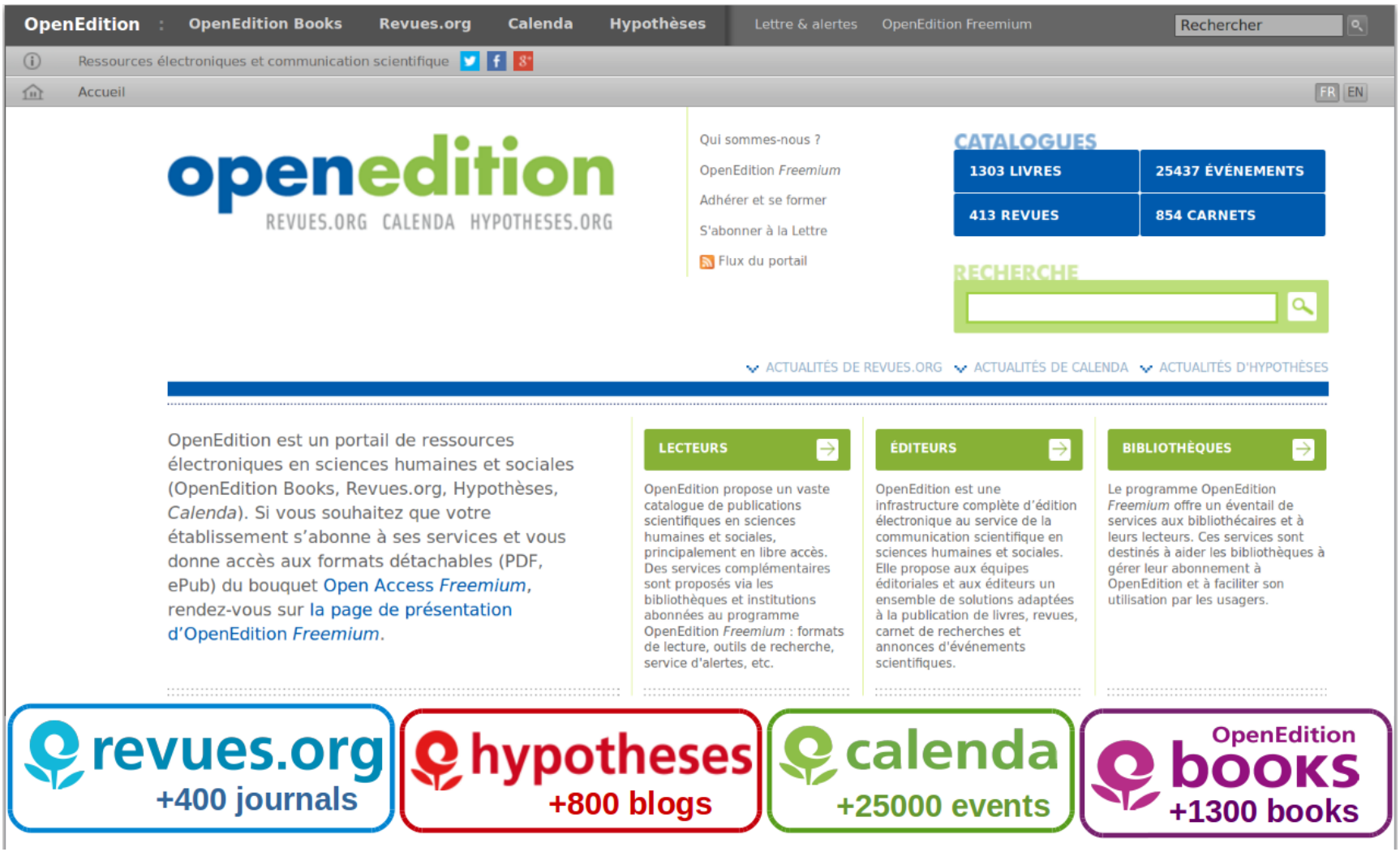
The volume shows how advances in the linguistic theory of event predicates, which have spawned Davidsonian and neo-Davidsonian notions of event arguments, in addition to 'event structure' frameworks and mereological models for the eventuality domain, have sidelined research on specific sets of entailments that support a typology of event predicates. Addressing this imbalance in the literature, the work also presents evidence indicating a more complex role for scalar structures than currently assumed. It will enrich the work of semanticists, psycholinguists, and syntacticians with a decompositional approach to verb phrase structure.

Objectif

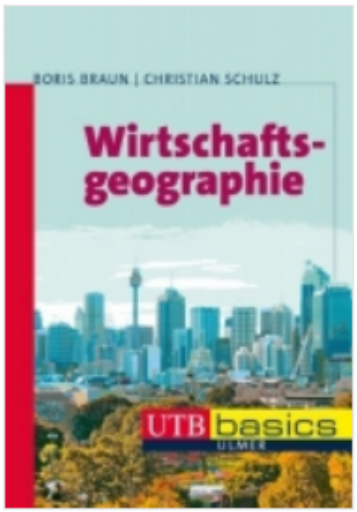
- Collecte de deux types de Comptes Rendus (CR):
  - Longs CRs des livres scientifiques écrits par des relecteurs experts dans des revues scientifiques;
  - Courts CRs trouvés dans le Web ou dans les réseaux sociaux.
- Mise en relation des CRs avec les livres correspondants en utilisant BILBO [1], un outil d'extraction et d'annotation automatique des réféfrences bibliographiques. (Développé par OpenEdition)

[1] Kim Y.M., Bellot P., Faath E., Dacos M. (2012b). Annotated Bibliographical Reference Corpora in Digital Humanities. 8th international conference on Language Resources and Evaluation (LREC)

Description



Compte Rendu de lecture - Revues.org



Full size image
Credits : © UTB GmbH

To come straight to the point, the new textbook *Wirtschaftsgeographie* [Economic Geography] by Boris Braun and Christian Schulz is a highly recommendable read. In pursuit of its goal to outline "*a very dynamic and multifarious sub-domain of geography*" [author's translation] (p. 249), this book is mainly directed at current and prospective students of the field of economic geography (p. 6). It keeps both these promises in quite an appealing manner. Alongside an insightful introduction to the myriad of theoretical structures applied in economic geography, its strength lies mainly in its engaging and stylistically coherent review as well as in the systematic cross-linkage of major topics. Valuable didactic tricks – such as info boxes giving a quick overview at the beginning of each chapter, additional text boxes that provide examples to underline the theoretical structures discussed, highlighted keywords in the body of the text, a rich stock of illustrations and charts, as well as review questions and recommendations for further reading – make this book a helpful companion for students. Topics are successfully cross-linked through clearly arranged and colour coded references of lexical precision as well as by repeatedly addressing, relating, and embedding major theoretical approaches from a variety of perception angles. The book embraces the exacting standards of both the authors, who explain them as follows: "*Currently relevant basic assumptions, theories, and models of economic geography, their differences as well as their manifold linkages shall be presented more vividly than in comparable textbooks*" [author's translation] (p. 6).

Corpus d'apprentissage et de test

Tous les documents ont été pré-classés en deux classes (CR et Non-CR). –> *Les documents ont été manuellement annotés.*

- Revues.org : (documents en français, format XML-TEI):
- Web: intérroger le moteur de recherche Google (concaténation du titre du livre avec tous ses auteurs).
  - Apprentissage : 358 CR et 294 Non-CR.
  - Test : 235 CR et 190 Non-CR.

Quels descripteurs pour une classification efficace ?

1. **Approche de "Sac de mots"** :
   - Texte = ensemble de mots.
   - Descripteur de texte = vecteur de poids.
   - Poids = occurrence d'un mot dans le texte.
   - Exclure toute analyse grammaticale et notion de distance entre les mots.

PB: espace vectoriel très large ($>$ 100 000 mots)

Quels descripteurs pour une classification efficace ?

2. Approche de sélection des caractéristiques :

Trouver un sous-ensemble "*pertinent*" de caractéristiques parmi celles de l'ensemble de départ. Pourquoi?

- Suppression des caractériqtiques redondantes –> simplifier la représentation des documents.
- Les caractéristiques non-nécessaires rajoutent du bruit au processus de prédiction
- Gain en rapidité (temps de prédiction pour chaque document)

Quels descripteurs pour une classification efficace ?

2. Approche de sélection des caractéristiques :

RFE-SVM (*Recursive Feature Elimination - Support Vector Machine*) –> méthode d'élimination progressive descendante (*Backward*)

- Plusieurs itérations sur l'ensemble des caractéristiques;
- Calculer le poids des caractéristiques et les ordonner (SVM);
- Supprimer progressivement dans chaque itération les caractéristiques de faible poids.

Quels descripteurs pour une classification efficace ?

3. **Approche basée sur la répartition des entités nommées** :
   - Analyse linguistique et statistique du corpus –> déterminer les caractéristiques communes entre les CR et les Non-CR
     - Titre des CR –> la référence bibliographique complète ou incomplète (titre, auteur, date)
     - Les documents Non-CR –> présense de la partie "*bibliographie*" à la fin du texte.
     - Les documents Non-CR –> présense des références implicites dans le texte (auteur, date).

Quels descripteurs pour une classification efficace ?

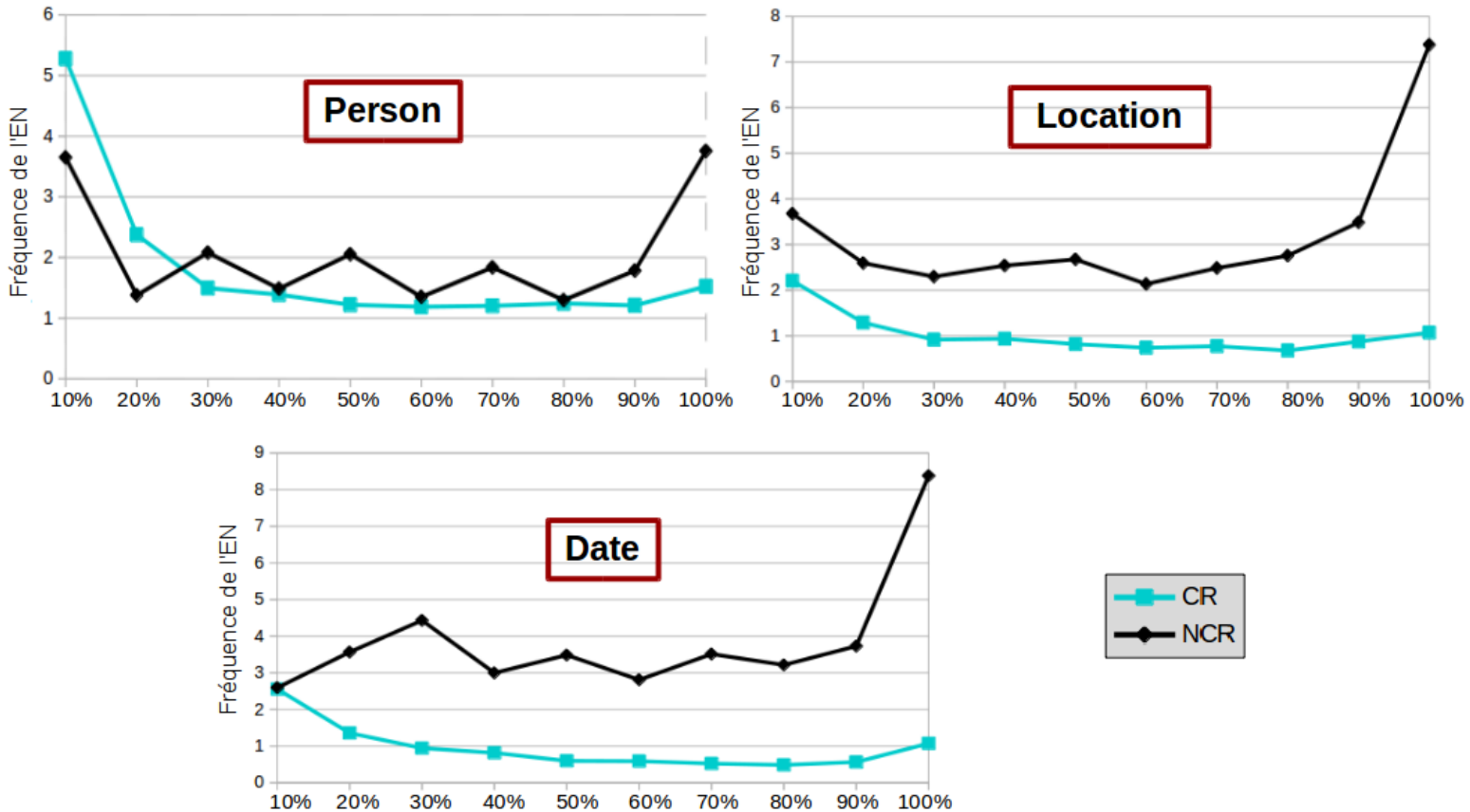
3. **Approche basée sur la répartition des entités nommées** :
   - Annoter des documents avec un détecteur d'entités nommées (TagEN [2])
     - Person
     - Date
     - Location
   - Diviser le texte de chaque document en 10 parties.
   - Calculer la répartiton des 3 EN dans chacune des parties.

[2] Poibeau T. «The multilingual named entity recognition framework». In EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, Morristown, NJ, USA, 2003, p. 155–158.

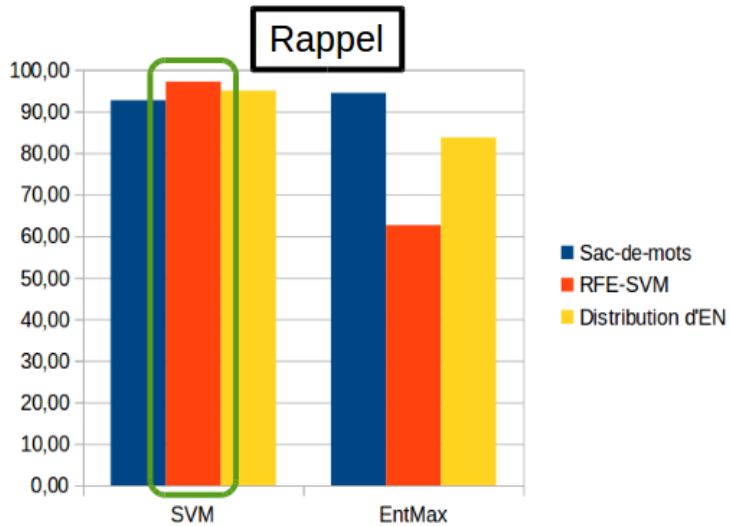
Quels descripteurs pour une classification efficace ?

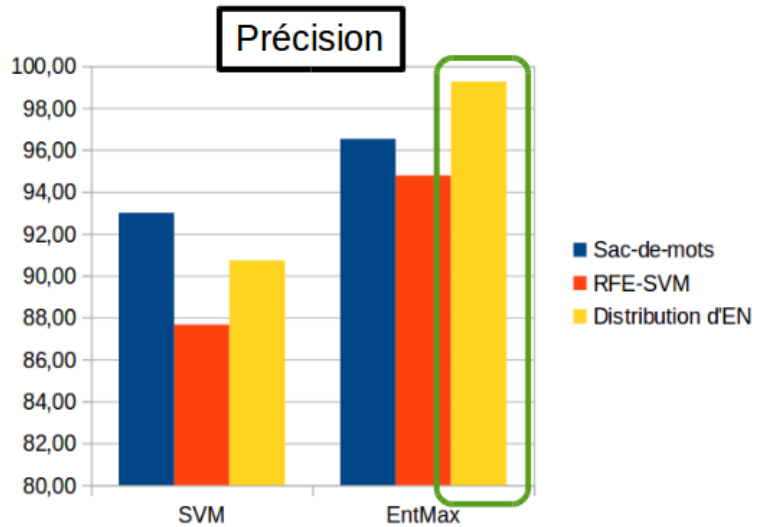
3. Approche basée sur la répartition des entités nommées :

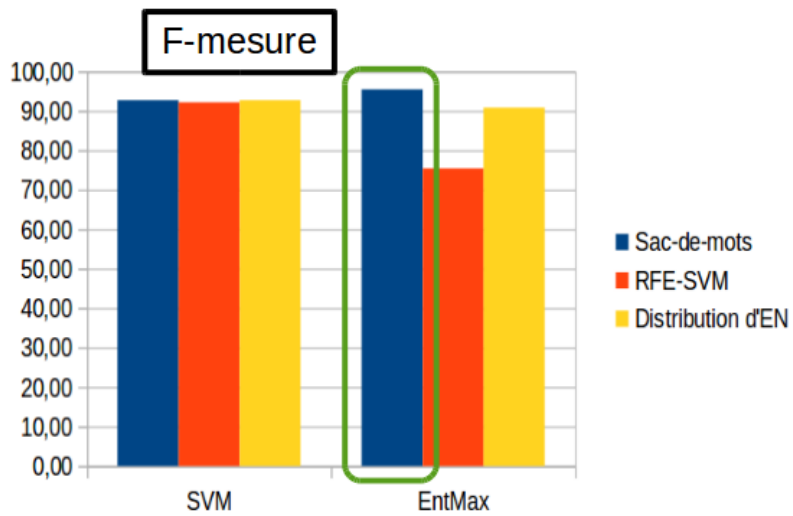
Figure : La distribution des Entités Nommées dans les deux classes (CR et Non-CR)



Résultats de la classification des documents en CR et Non-CR







Perspectives

- Utilisation d'autres méthodes de sélection de caractéristiques (TF-IDF, Z-score, ...)
- Exploiter l'analyse de polarité pour améliorer les performances de classification.
- Classification des blogs scientifiques pour identifier les CR.
- Extraire la référence du livre dans un CR en utilisant BILBO.

Conclusion

- Deux collections de Compte Rendu de lecture: les platformes d'OpenEdition (XML-TEI) et le Web (HTML et Java Scripts).
- Nouvelles méthodes de classification en genre basées sur les approches de classification thématique.
- Utilisation de la distribution des Entités Nommées dans le texte comme caracteristique discriminante.
- Efficacité des méthodes de classification supervisée pour l'identification des textes porteur d'opinion.

The collection can be freely downloaded (OAI-PMH). Please contact us at: lab@openedition.org

**** Merci pour votre attention ****