Comparaison de méthodes sémantique et asémantique pour la catégorisation automatique de documents

Romaric Boley, M.S.I.
École de bibliothéconomie et des sciences de l'information
Université de Montréal
✉ romaric.boley@umontreal.ca

7e défi fouille de textes (DEFT 2011)
Montpellier, 1er juillet 2011

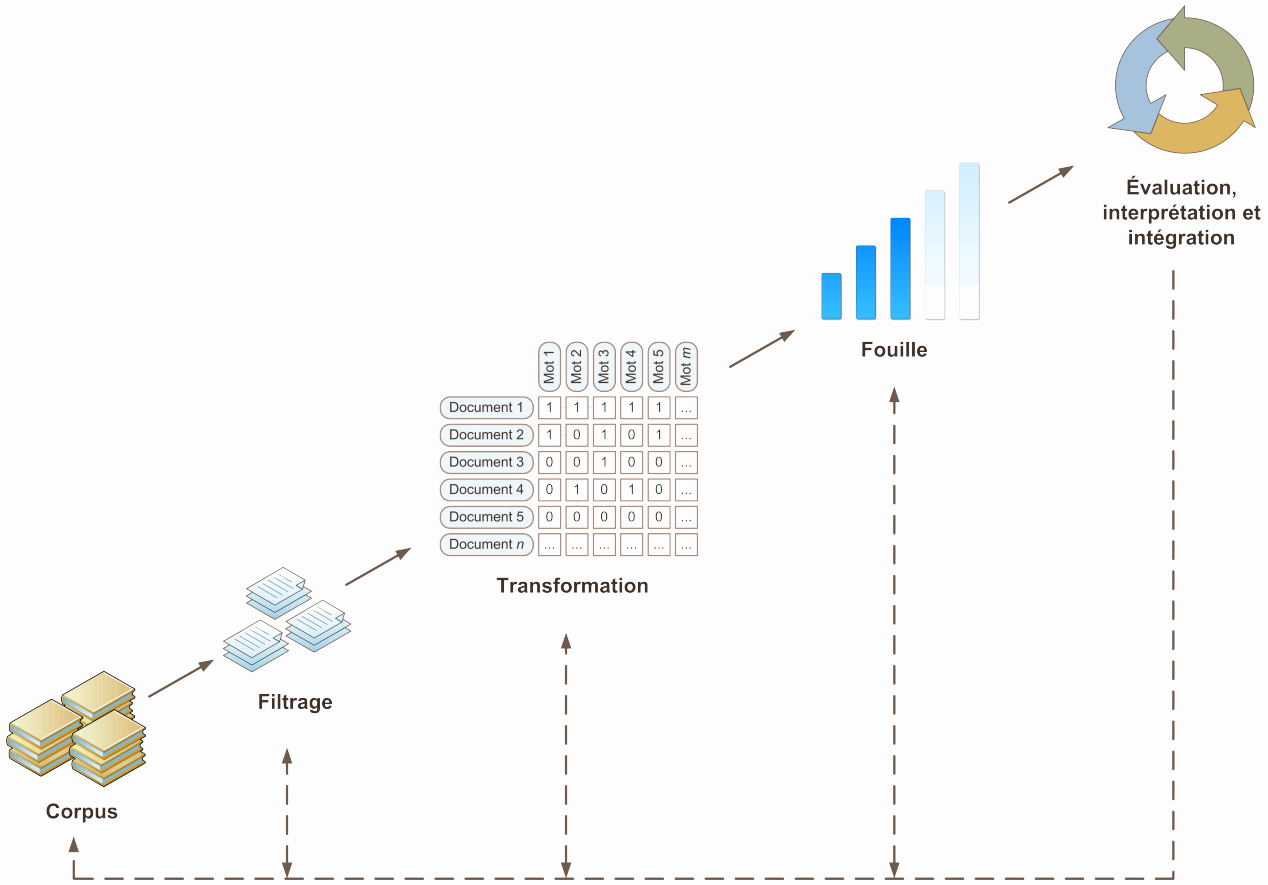
Plan

- Introduction
- Méthodologie
- Tâche 1 - Résultats
  - Phase d’apprentissage
  - Phase de test
- Tâche 2 : Résultats
  - Phase d’apprentissage
  - Phase de test
- Conclusion

Introduction

- Objectif : évaluer l’impact d’une variation au niveau de la nature des traits discriminants utilisés pour décrire des documents dans un contexte de catégorisation automatique
  - Utilisation d’algorithmes de catégorisation classiques
  - Dans la poursuit des travaux de Forest dans le cadre de DEFT 2009

Méthodologie



Méthodologie

- Approche sémantique vs asémantique
  - Sémantique :
    - Extraction du lexique avec application d'un anti-dictionnaire pour ne garder que les mots porteurs de sens
    - Pondération des termes selon le TF*IDF
  - Asémantique :
    - Mots non signifiants
      - Extraction du lexique uniquement des mots non porteurs de sens (sélection des termes présents uniquement dans l'anti-dictionnaire)
    - Longueur des phrases
      - Extraction des longueurs des phrases (nombre de *tokens*) sans filtrage du lexique
- Aucune correction du corpus

Tâche 1 – Variation diachronique

- Paramètres définis lors de la phase d’apprentissage
  - Pondération des attributs : Chi2
  - Méthode d’échantillonnage : Retrait d’un cas (*leave one out*)
  - Algorithme : Classifieur Bayésien naïf

Tâche 1 – Variation diachronique

- Résultats de la phase d'apprentissage (500 mots)

Stratégie	Nombre d'attributs	Score
Longueur des phrases	154	7,23%
Mots non porteurs de sens	505	9,75%
Mots porteurs de sens	103897	39,44%

Tableau 1 : Apprentissage pour le corpus contenant des extraits de 500 mots

Tâche 1 – Variation diachronique

- Résultats de la phase d'apprentissage (300 mots)

Stratégie	Nombre d'attributs	Score
Longueur des phrases	137	7,09%
Mots non porteurs de sens	495	8,40%
Mots porteurs de sens	75116	32,78%

Tableau 2 : Apprentissage pour le corpus contenant des extraits de 300 mots

Tâche 1 – Variation diachronique

- Résultats de la phase de test (500 mots)

Test	Score
1-Longueur des phrases	6,2%
2-Mots non porteurs de sens	7,3%
3-Mots porteurs de sens	6,1%

Tableau 3 : Résultats pour la variation diachronique sur les extraits de 500 mots

Tâche 2 – Appariement résumé / article

- Paramètres définis lors de la phase d'apprentissage
  - Pondération des attributs : TFxIDF
  - Variation du nombre de traits discriminants : entre 100 et 1000
  - Méthode d'échantillonnage : Retrait d'un cas (*leave one out*)
  - Mesure de similarité : cosinus (requête / document, Salton 1988)

Tâche 2 – Appariement résumé / article

- Résultats de la phase d'apprentissage (articles complets)

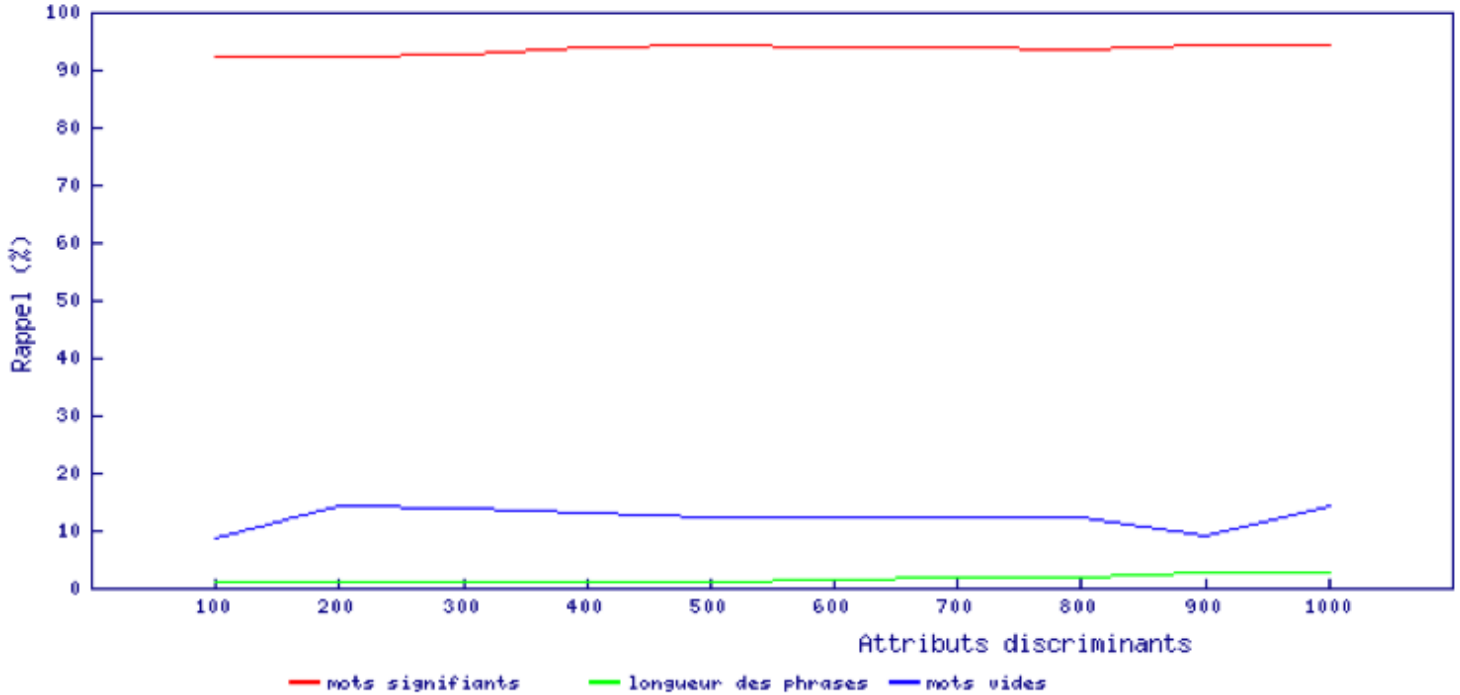


Figure 2: Performances de l'appariement, résumé/article complet, suivant les stratégies adoptées

Tâche 2 – Appariement résumé / article

- Résultats de la phase d'apprentissage (articles incomplets)

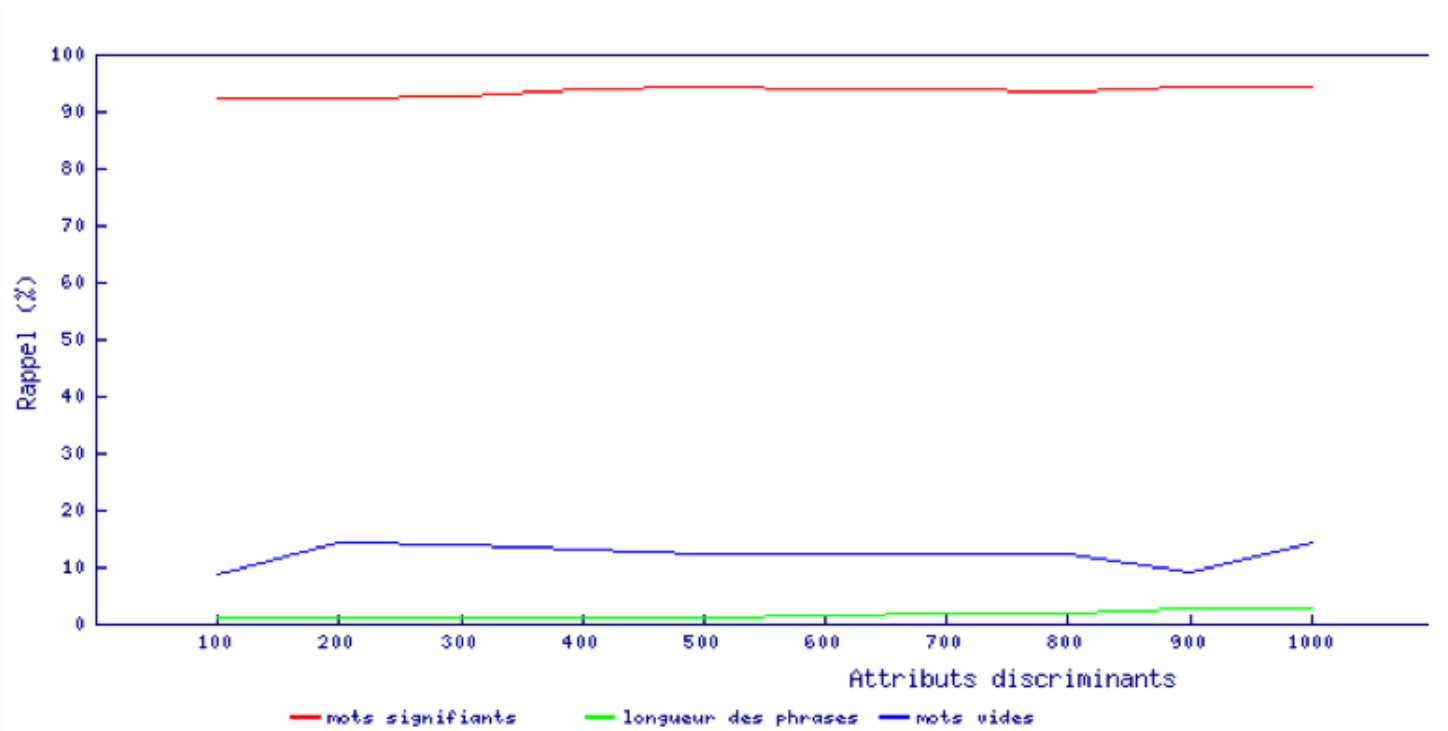


Figure 1: Performances de l'appariement, résumé/article sans introduction ni conclusion, suivant les stratégies adoptées

Tâche 2 – Appariement résumé / article

- Résultats de la phase de test (articles complets)

Test	Nombre d'attributs	Score global
1	600	98%
2	800	98,5%
3	1000	98%

Tableau 4 : Résultats pour l'appariement résumé / article complet

Tâche 2 – Appariement résumé / article

- Résultats de la phase de test (articles incomplets)

Test	Nombre d'attributs	Score global
1	600	98%
2	800	98,5%
3	1000	98%

Tableau 4 : Résultats pour l'appariement résumé / article complet

Conclusion

- Classifieurs victimes de surapprentissage
- Méthode de validation *leave-one-out* :
- La correction des termes du corpus semble essentielle
  - En se basant sur les les réformes orthographiques
- Stratégies asémantiques plus robustes ?