Expérimentations autour des espaces sémantiques hybrides.

Adil El Ghali
avec les contributions de
Kaoutar El Ghali, Sylvain Baron, Louis-Gabriel Pouillot

IBM CAS France & Lutin UserLab

July 1, 2011

Introduction

- *Alida*, le système développé pour le DEFT'09 et DEFT'10 a servi de base pour le DEFT'11
- Notre but pour cet édition ťait de tester certaines hypothèses:
  - L'apport de l'utilisation des informations liées au bruit d'OCR.
  - La stabilité des espaces sémantiques lors de l'ajout d'information.
  - Les performances du système en réduisant la taille de l'espace sémantique.

Alida

- Un modèle pour la catégorisation de texte se basant sur :
  - Capture de la sémantique latente des documents
  - Utilise les espaces sémantiques comme modèle de reprèsentation de la mémoire
  - Les catégories sont divisées en sous-catégories (*cibles*) pour prendre en compte les particularité de certains épisodes

Alida: Construction des espaces sémantiques

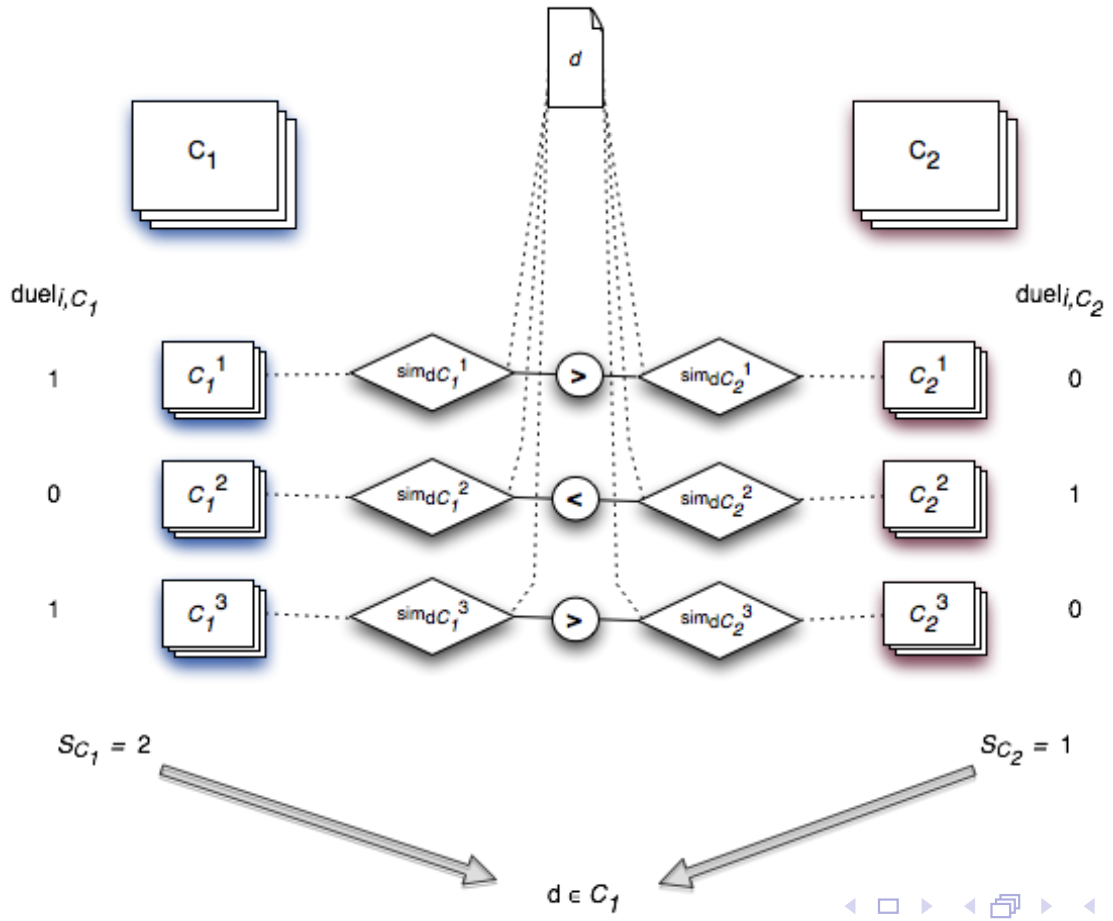
- Les espaces sémantiques sont construits en utilisant *Random Indexing*
  - Créer une matrice $A(d \times N)$, contenant des vecteurs-index:
    - d est le nombre de documents dans le corpus
    - N le nombre de dimensions ($N > 1000$)
    - les vecteurs-index sont des vecteurs creux générés aléatoirement: $[0, ..., 0, +1, 0, ..., 0, -1, ...]$
  - Créer une matrice $B(t \times N)$ contenant des vecteurs-term:
    - t est le nombre de termes composant le corpus
    - Au départ la matrice B est initialisée avec des valeurs nulles.
  - Pour tout document d dans le corpus, à chaque fois que le terme t apparaît dans d
    - On accumule le vecteur-index de d au vecteur-terme de t
  - A la fin su processus, les termes qui sont apparu dans les mêmes contextes auront des vecteurs similaires.

Alida: Construction des cibles

- Les cibles sont des prototypes de la catégorie, ils sont obtenues par une partition de l'ensemble des vecteurs représentant les documents d'une catégorie
  - Etant donnée une catégorie $C = \{d_0, ..., d_n\}$
  - On définit $C' = \{d'_0, ..., d'_n\}$ avec $\forall i, ji < j \Rightarrow sim(d'_i, C) < sim(d'_j, C)$
  - Pour obtenir t cibles, il suffit de découper C' en t sous ensembles de même cardinalité $C' = [\underbrace{d'_0, ..., d'_{n_1}}_{C^1}, ..., \underbrace{d'_{n_{t-1}}, ..., d'_n}_{C^t}]$

Alida: Attribution de catégories

- On calcule la similarité d'un document à catégoriser avec l'ensemble des cibles de chaque catégorie, et on les compare celles des cibles des autres de même rang

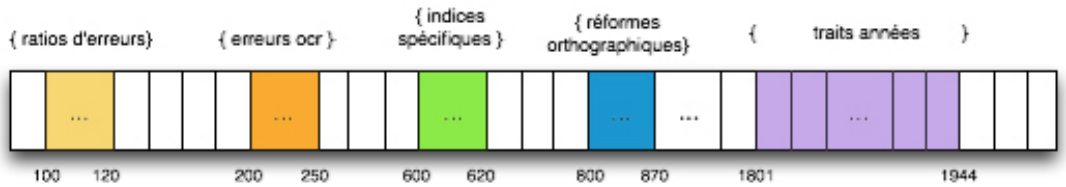


Nettoyage du corpus

- Le corpus diachronique qui nous a été fourni était issu de l'OCRisation de documents
- Plusieurs stratégies de corrections, on été effectuée
  - une correction manuelle d'échantillons du corpus et une propagation de ces corrections sur la totalité du corpus
  - le repérage d'erreurs systématiques et écriture d'une base de régles les prenant en compte dans le correcteur orthographique `hunspell`

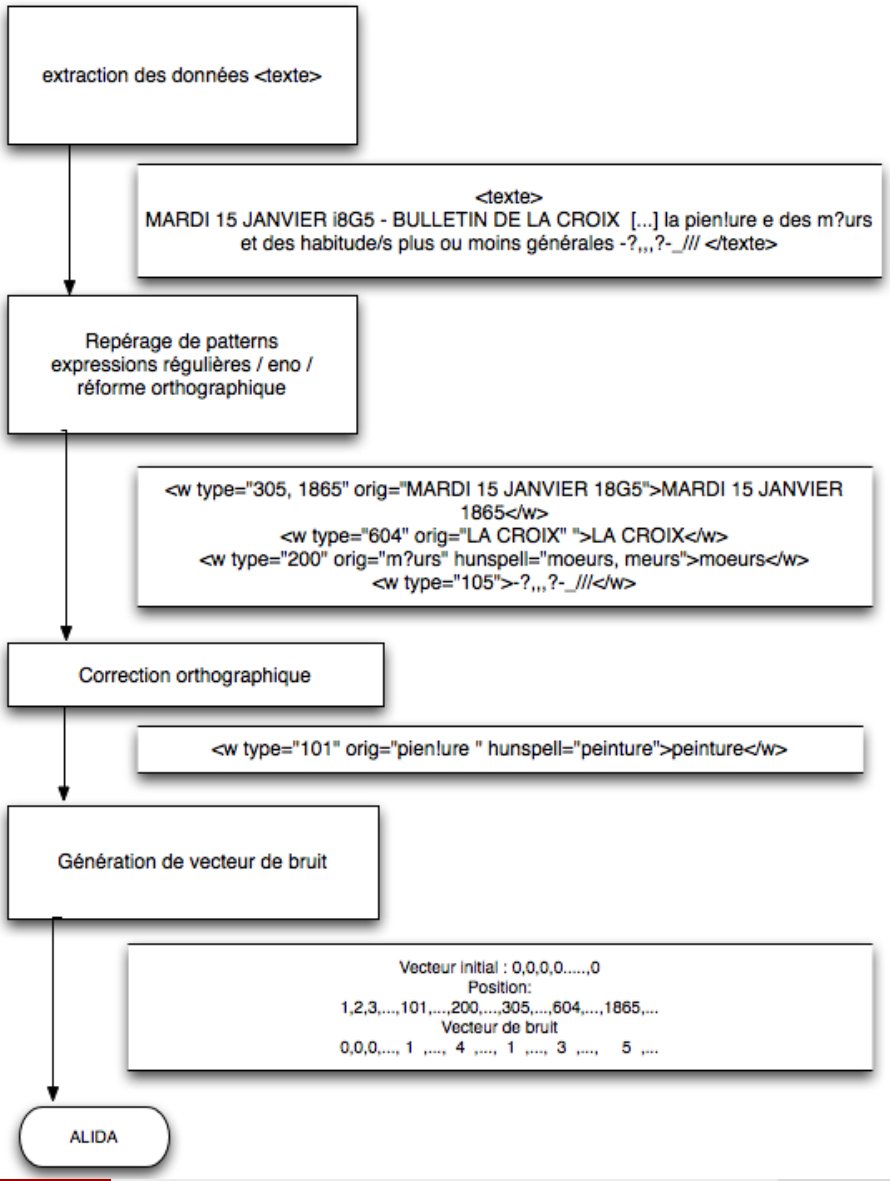
Vecteurs de corrections

- Taux d'erreurs
- Réformes orthographiques
- Erreurs d'OCR
- Traits d'années



- Insertion de documents *thématiques* pour les années recherchées

Chaîne de traitement



Exécutions soumises

- Application d'*Alida* avec les 144 catégories
- Utilisation du vecteur Wikipedia de l'année comme représentant de la catégories

Exécution	Description	F-score
#1	Alida (wikipedia)	0.098
#2	Alida doc + bruit	0.108
#3	Alida bruit	0.100
#1	Alida corpus bruité	0.113
#2	Alida corpus corrigé + vecteur de bruit	0.117
#3	Alida (wikipedia)	0.081

Appariements

- L'espace sémantique contient deux types de documents: Article, Résumé
- L'appariement entre un résumé R_i et un article A se fait par la projection du Vecteur V_{R_i} sur le plan défini par l'ensemble des vecteurs des articles.
- Pas d'apprentissage.

Exécutions soumises

- On a fait varié la dimension des espaces sémantiques

Exécution	Dimension	F-score	Temps d'exec
#1	1000	0.970	30s
#2	500	0.965	10s
#3	200	0.934	4s
#1	1000	0.919	25s
#2	500	0.898	9s
#3	200	0.883	3s

Conclusions

- Les résultats sont assez décevants.
  - L'ajout d'informations dans l'espace (pages wikipedia) pose problème, ces pages se retrouvent dans une zone à part dans l'espace.
  - Les corrections de corpus permettent d'améliorer les résultats
- La tâche 2 montre qu'on peut aller vite.

Conclusion

C'est FUN

Malgré les nuits blanches et les grandes doses de caféine