LSVMA : au plus deux composants pour apparier des résumés à des articles

Yves Bestgen

F.R.S-FNRS et UCL / IPSY / CECL
Louvain-la-Neuve - Belgique

Plan

- Le problème
- L'approche
- Analyses et résultats
- Utilité des composants
- Conclusion

L'approche

L*S*V*M*A : Trois composants

LSVMA : L'analyse sémantique latente

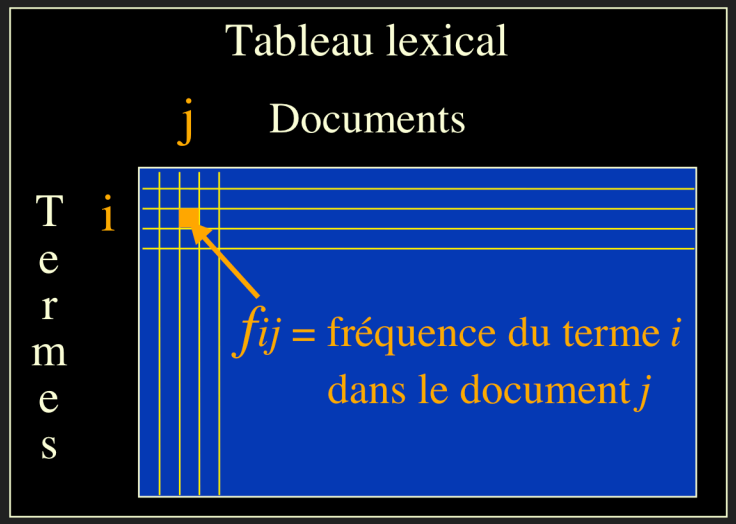
LSVMA : L'analyse sémantique latente

- Pourquoi?
  - Foltz, Britt et Perfetti (in Foltz, 1996)
    - Quels textes ont le plus influencé un résumé?
    - Apparier les phrases du résumé et celles des textes
  - Evaluation automatique de résumés
    - Comparer les résumés à un document de référence
    - Le plus similaire est le meilleur

(Kintsch et al., 2000 ; Olmos et al., 2009)

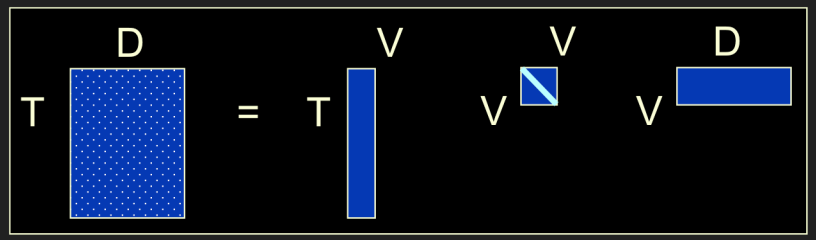
LSVMA : L'analyse sémantique latente

- Comment?
  - Corpus →



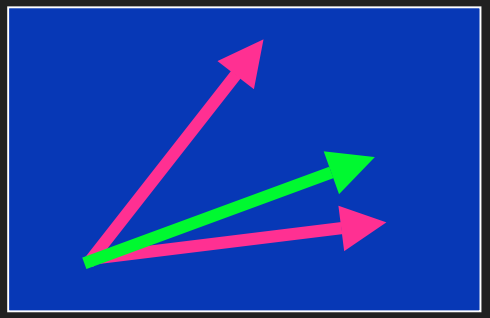
LSVMA : L'analyse sémantique latente

- Comment?
  - Corpus → Tableau lexical
  - Décomposition en valeurs singulières et réduction
    - Espace sémantique



LSVMA : L'analyse sémantique latente

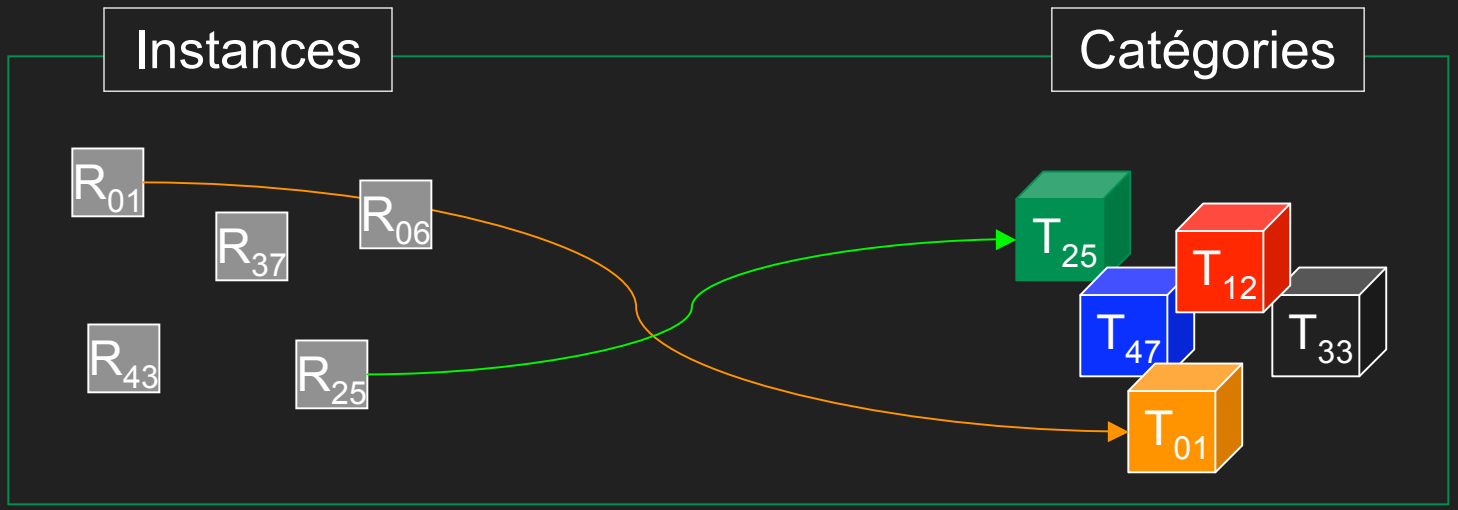
- Comment?
  - Corpus → Tableau lexical
  - Décomposition en valeurs singulières et réduction
  - Proximité entre "résumé" et "document de référence"



LSVMA : Machine à support vectoriel

LSVMA : Machine à support vectoriel

- Pourquoi?
  - Un problème de catégorisation de textes



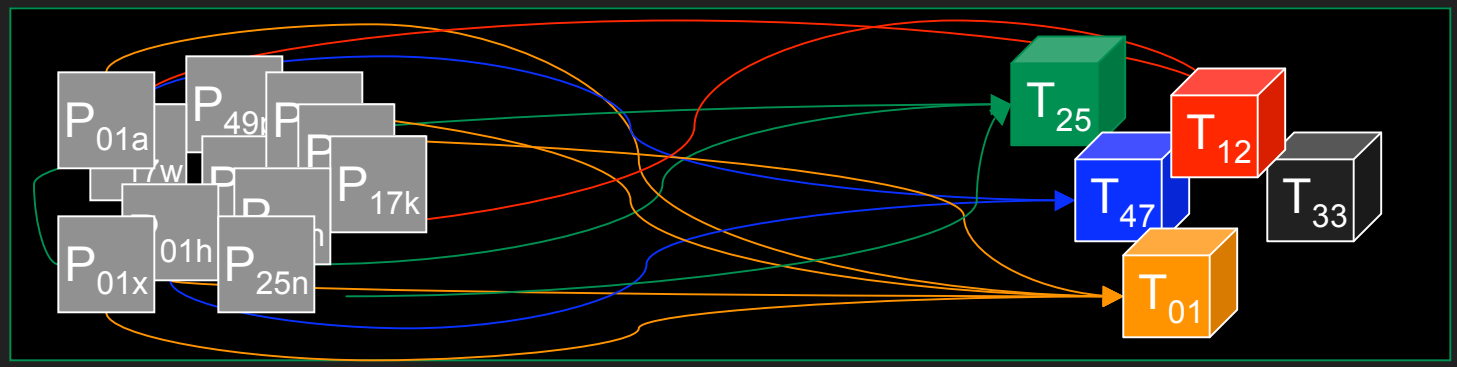
  - Classifieur SVM très efficace, y compris à partir de LSA

LSVMA : Machine à support vectoriel

- Comment?
  - Classer chaque résumé dans le texte correspondant sur la base des vecteurs de LSA
  - Mais sur quelle base apprendre?

LSVMA : Machine à support vectoriel

- Comment?
  - Classer chaque résumé dans le texte correspondant sur la base des vecteurs de LSA
  - Mais sur quelle base apprendre?
    - Classer les paragraphes dans les textes



- Multicatégorielle : *SVM*^{multiclass} (Joachims et al., 2009)

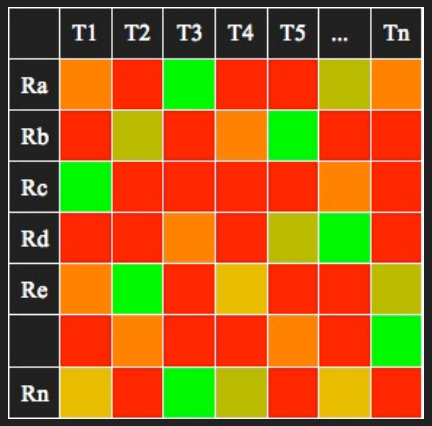
LSVMA : Affectation au Meilleur d'Abord

LSVMA : Affectation au Meilleur d'Abord

- Pourquoi?
  - Un cas particulier de catégorisation de textes
    - Relation biunivoque entre résumés et textes
      - Non prise en compte par la SVM

LSVMA : Affectation au Meilleur d'Abord

- Comment?
  - Un problème d'affectation ou d'appariement



(Rubin, 1973 ; Rosenbaum, 2010)

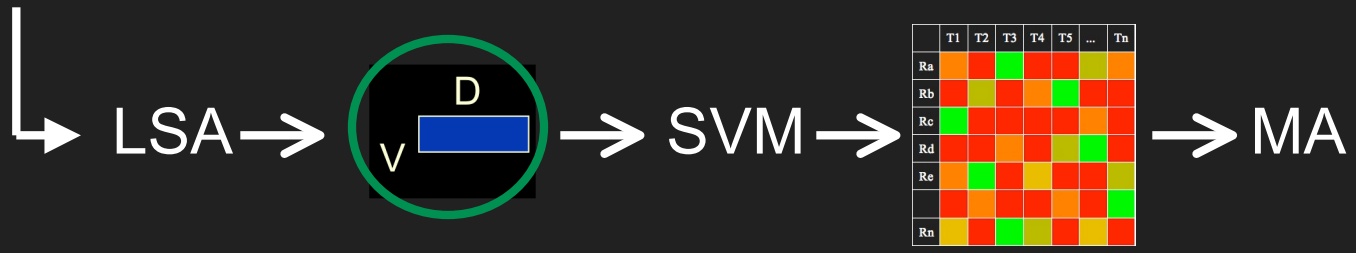
  - Solution simple
    - Technique du plus proche disponible en commençant par le meilleur couple Résumé-Texte

LSVMA : Implémentation

- Prétraitements (principaux)
  - Lemmatisation (Treetagger : Schmid, 1994)
  - Textes segmentés en paragraphes de minimum 75 mots
  - Une matrice par revue : Termes x [paragraphes+résumés]

LSVMA : Implémentation

- Prétraitements (principaux)
  - Lemmatisation (Treetagger : Schmid, 1994)
  - Textes segmentés en paragraphes de minimum 75 mots
  - Une matrice par revue : Termes x [paragraphes+résumés]



LSVMA : Implémentation

- Prétraitements (principaux)
  - Lemmatisation (Treetagger : Schmid, 1994)
  - Textes segmentés en paragraphes de minimum 75 mots
  - Une matrice par revue : Termes x [paragraphes+résumés]

- Paramètre pour l'optimisation
  - Nombre de vecteurs de l'espace sémantique
    - 300, 200 et 100

Analyses et résultats

Apprentissage : Classifications des paragraphes

Piste	Nvec	ANT	MET	SCI	INT	LIT
1	100					
	200					
	300					
2	100					
	200					
	300					

Classifications correctes (%)

Apprentissage : Classifications des paragraphes

Piste	Nvec	ANT	MET	SCI	INT	LIT
1	100	93	80	94	95	95
	200	97	90	97	97	98
	300	98	94	98	98	99
2	100	93	81	93	94	94
	200	98	91	96	97	98
	300	99	94	98	98	99

Classifications correctes (%)

Apprentissage : Classifications des paragraphes

Piste	Nvec	ANT	MET	SCI	INT	LIT
1	100	93	80	94	95	95
	200	97	90	97	97	98
	300	98	94	98	98	99
2	100	93	81	93	94	94
	200	98	91	96	97	98
	300	99	94	98	98	99

Classifications correctes (%)

Appariement : Classifications des résumés

Piste	Nvec	ANT	MET	SCI	INT	LIT
1	100	100	90	100	100	100
	200	100	100	100	100	100
	300	100	100	100	100	100
2	100	100	90	100	100	100
	200	100	97	100	100	100
	300	100	100	100	100	100

Classifications correctes (%)

Appariement : phase de test

- Les 3 valeurs de Nvec ont donné lieu au même appariement
  - Une seule soumission par piste : 100%

LSVMA : Utilité des trois composants?

LSVMA : Utilité des trois composants?

Adapté de http://reviews.goldenagecartoons.com/2010/mighty/

LSVMA : Utilité des trois composants?

- SVM est le noyau de l'approche
- LSA ?

- MA ?

LSVMA : Utilité des trois composants?

- SVM est le noyau de l'approche
- LSA ?
  - Raison d'être : synonymie (au sens large), mais ici?
    - SVM sur matrice Termes x Documents
      - Performances aussi bonnes qu'avec LSA
    - LSA non nécessaire
- MA ?

LSVMA : Utilité des trois composants?

- SVM est le noyau de l'approche
- LSA ?
  - Raison d'être : synonymie (au sens large), mais ici?
    - SVM sur matrice Termes x Documents
      - Performances aussi bonnes qu'avec LSA
    - LSA non nécessaire
- MA ?
  - Sans MA, performances moins bonnes
    - Semble nécessaire...

Conclusion

- **LSVMA** : LSA + SVM + MA
- Pour la tâche, LSA n'est pas nécessaire
  - Conclusion généralisable à l'évaluation d'*hétéro-résumé*?

(Schnedecker, 2001)

Conclusion

- **LSVMA** : LSA + SVM + MA
- Pour la tâche, LSA n'est pas nécessaire
  - Conclusion généralisable à l'évaluation d'*hétéro-résumé*? (Schnedecker, 2001)
- Limitation de l'approche
  - Pourquoi de "moins bonnes" performances pour *Meta*?
    - Pour l'appariement, mais aussi pour l'apprentissage

Apprentissage : Classifications des paragraphes

Piste	Nvec	ANT	MET	SCI	INT	LIT
1	100	93	80	94	95	95
	200	97	90	97	97	98
	300	98	94	98	98	99
2	100	93	81	93	94	94
	200	98	91	96	97	98
	300	99	94	98	98	99

Classifications correctes (%)

Conclusion

- **LSVMA** : LSA + SVM + MA
- Pour la tâche, LSA n'est pas nécessaire
  - Conclusion généralisable à l'évaluation d'*hétéro-résumé*? (Schnedecker, 2001)
- Limitation de l'approche
  - Pourquoi de "moins bonnes" performances pour *Meta*?
    - Pour l'appariement, mais aussi pour l'apprentissage
      - La revue ou le format ?

Différence de format entre les revues

Revue des sciences de l'éducation

<titre>Introduction</titre>

En 1983, Viviane Isambert-Jamati, qui retraçait l'histoire des rapports entre les sciences sociales et « le ministère » dans le domaine de l'éducation, concluait à une ignorance réciproque (Berthelot, Forquin, Isambert-Jamati et Tanguy, 1984). En 2003, Franck Poupeau dénonce une [...]

Meta

Le modèle présenté ici [...] : il modélise l'activité
professionnelle standard [...] « traducteur ». Il est
centré sur l'ensemble [...] matériau traduit autonome

GENESE

Merci de votre attention

Références

Foltz (1996). Latent Semantic Analysis for text-based research. *Behavior Research Methods* 28, 197–202.

Joachims et al. (2009). Cutting-Plane training of structural SVMs, *Machine Learning* 77, 27–59.

Kintsch et al. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments* 8, 87–109.

Olmos et al. (2009). New algorithms assessing short summaries in expository texts using Latent Semantic Analysis. *Behavior Research Methods* 41, 944–950

Rosenbaum (2010). *Design of Observational Studies*. Springer.

Rubin (1973). Matching to remove bias in observational studies, *Biometrics* 29, 159-183.

Schmid (1994). Probabilistic part-of-speech tagging using decision trees. *1st NMLP Conf.*, 44–49.

Schnedecker (2001). *Lire, comprendre, rédiger des textes théoriques*. De Boeck.

Analyses et résultats

- Matériel pour le développement
  - 5 revues
    - ± 8 000 termes différents par revue
    - En moyenne, de 22 à 44 paragraphes par texte
- Matériel pour le test
  - 6 revues

LSA : Pondération log-entropie

- Pondération locale
  - Réduit l'impact des mots très fréquents dans un document

$$f_{ij}' = \frac{\log(f_{ij} + 1)}{-\sum_j \frac{f_{ij}}{\sum_j f_{ij}} \log(\frac{f_{ij}}{\sum_j f_{ij}})}$$

- Pondération globale
  - Réduit l'impact des mots peu informatifs
    (Fréquence similaire dans tous les documents)