Comment faire pour que l'opinion forgée à la sortie des urnes soit la bonne ?
Application au défi DEFT 2007

Juan Manuel Torres-Moreno
Marc El-Bèze, Fréderic Béchet, Nathalie Camelin

3 juillet 2007

Laboratoire Informatique d'Avignon/EA 931
Université d'Avignon et des Pays de Vaucluse

Stratégie

- Une dizaine d'approches numériques basées sur l'apprentissage automatique
  - But : reproduire la règle d'association texte/opinion à partir d'un corpus étiqueté
- Méthode
  - Déployer de nombreux systèmes avec de multiples représentations des données
    - But : plutôt que de régler très précisément un seul système sur le corpus d'apprentissage (avec risque de sur-apprentissage), entraîner un grand nombre de systèmes (certains sans adaptation)
  - Apprentissage sur corpus avec « 5-fold cross validation »
    - But : toujours éviter le « sur-apprentissage » des modèles
  - Fusionner les résultats des différents systèmes
    - Vote simple pour DEFT'07

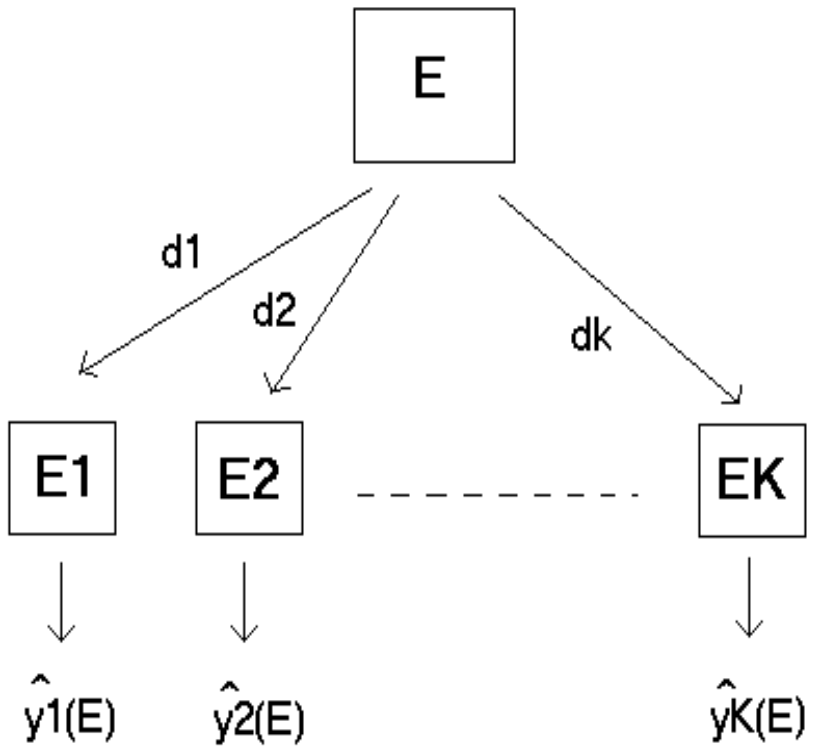
Systèmes utilisés

- Classifieurs « sortis de leur boîte »
  - BoosTexter : AdaBoost (Schapire & Singer)
  - SVM-Torch : Support-Vector-Machines (Vapnick)
  - Timble : *K*-plus proches voisins (Daelemans & Van den Bosch)
  - LIA_SCT : Arbres de classification sémantiques (Kuhn & de Mori)
- Classifieurs adaptés pour DEFT'07
  - LIA_JUAN : Modèle de probabilités n-grammes avec/sans lemmatisation
  - LIA_MARC : Modélisation probabiliste discriminante

BoosTexter : Algorithme de boosting

Améliorer la précision des règles de classification en combinant plusieurs hypothèses « faibles »

- Re-pondération répétitive des exemples du jeu d'entraînement
- Ré-exécution sur ces données re pondérées
- Focaliser sur les exemples les plus difficiles à classer

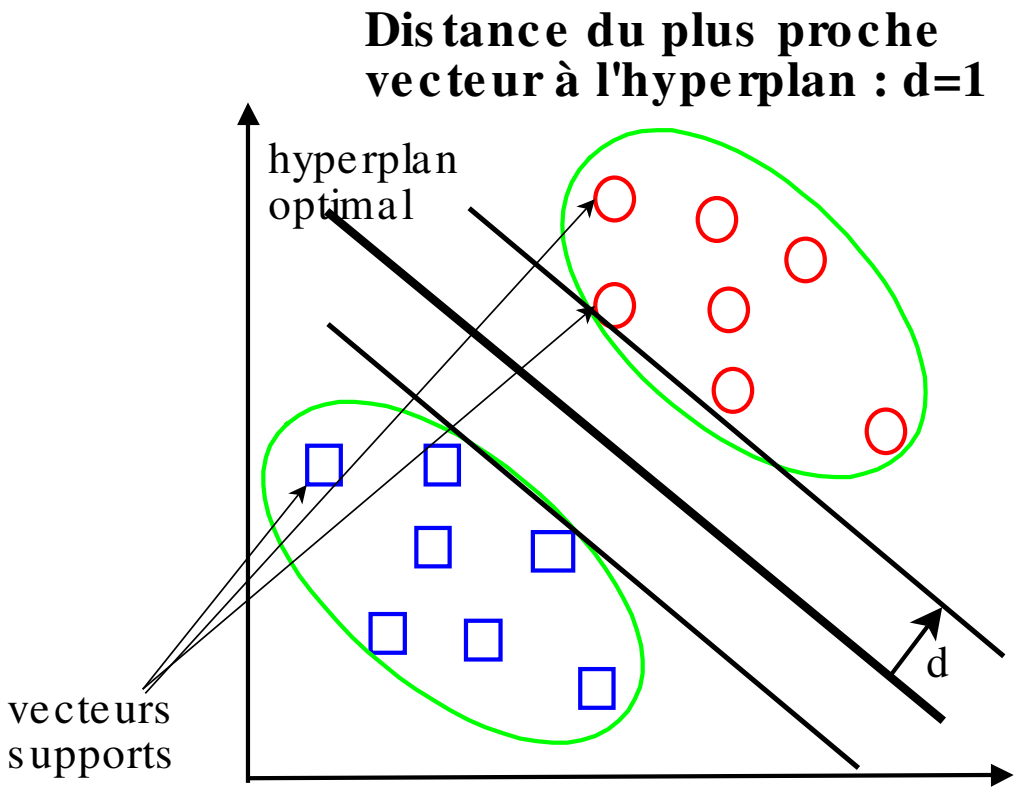


SVMTorch : Machines à support vectoriels

Ce classifieur à large marge découpe le problème de classification en 2 sous-problèmes:

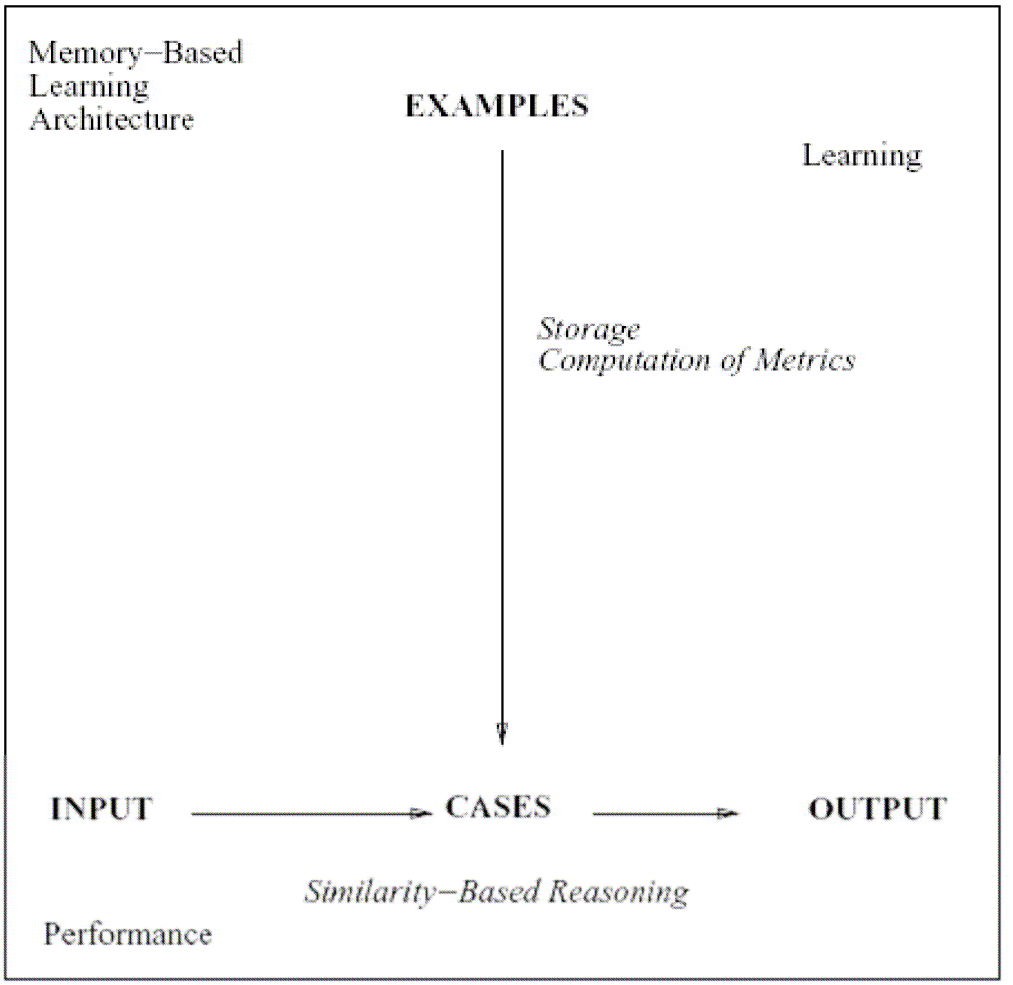
- trouver un espace où les données sont linéairement séparables

➢séparer les données grâce à un hyperplan maximisant la marge



TiMBL : Tilburg Memory-Based Learner

- Classification fondée sur les K-plus proche voisins
- Apprentissage : stockage d'exemples
- Classification : recherche de l'exemple le plus «proche» dans la base d'exemples grâce à diverses mesures de similarité



LIA_SCT : Arbres de décision sémantiques

Un arbre de décision où chaque question posée à un nœud est une expression régulière construite automatiquement

NODE	Expression régulière	Classe	Proba
27	+ par_exemple + excellent NMS +	2	0.8148
32	+ recherche + excellent +	2	0.5500
34	+ excellent + 3D +	2	0.6154
36	+ excellent + graphique +	2	0.5200
44	+ moche + sympathique +	1	0.8148
50	+ moche + un_peu +	0	0.6190
52	+ pourtant + moche +	0	0.6786
54	+ moche + assez +	0	0.8125
55	+ moche +	0	1.0000
66	+ nouveau + sympathique + plaisir + mode +	2	0.5000
67	+ nouveau + sympathique + plaisir +	1	0.4286
71	+ nouveau + offrir + sympathique +	1	0.5556
73	+ nouveau + aller + sympathique +	1	0.7632
145	+ catastrophique +	0	0.6897
152	+ assez + aucun +	0	0.6216
158	+ assez + maniabilité +	1	0.5217

LIA_JUAN

- Filtrage très legèr afin de capturer des tournures
  - Voix passive, formes interrogatives et exclamatives
- Aggregation de mots dans la même famille
  - Collocations + morphologique
- Utilisation d'un modèle d'uni-lemmes
- Probabilité d'appartenance à une classe:

$$P_t(w) \approx \prod_i \lambda_1 P_t(w_i) + \lambda_0 U_0$$

LIA-MARC

- **Du symbolique pour améliorer le numérique**
  - Lemmatisation (LIA_TAG) sauf exceptions
    - *« serait préférable », « vous auriez dû »*
  - Normalisation des textes (30 000 puis 35 000 règles)
    - Uniformisation des variantes graphiques
      - *Fautes d'orthographe, casse, quelques traductions, etc.*
- **Du numérique et du symbolique**
  - Agglutination par règles (30 000 puis 60 000) pour la plupart issues d'un calcul de collocation amélioré

LIA-MARC : Modèle

Processus de décision

$$\hat{t} = \underset{\mathbf{t}}{\mathbf{Argmax}}\, P(t) \times P(w \mid t) = \underset{\mathbf{t}}{\mathbf{Argmax}}\, P(t) \times P_t(w)$$

Second terme : combinaison classique de *ngram*

$$P_t(w) \approx \prod_i \lambda_3 \times P_t(w_i \mid w_{i-2} w_{i-1}) + \lambda_2 \times P_t(w_i \mid w_{i-1}) + \lambda_1 \times P_t(w_i) + \lambda_1 U_0$$

Estimation discriminante des paramètres : fréquences pondérées via un critère proche du critère de Gini

$$G(w,h) \approx \sum_t P^2{}_t(t \mid w,h)$$

Représentation des textes

- Tokens
  - Mots, Part-Of-Speech, Lemmes
  - *Mots seed* ***(Wilson et al., 2005)***
    - Ensemble de mots susceptibles d'être associés à l'expression d'une opinion
    - Liste manuelle + sélection automatique sur le corpus d'apprentissage avec BoosTexter
    - Lexique de 2000 seeds
- Textes
  - Suites de tokens
  - Sacs de tokens
  - Sacs de n-grammes de tokens

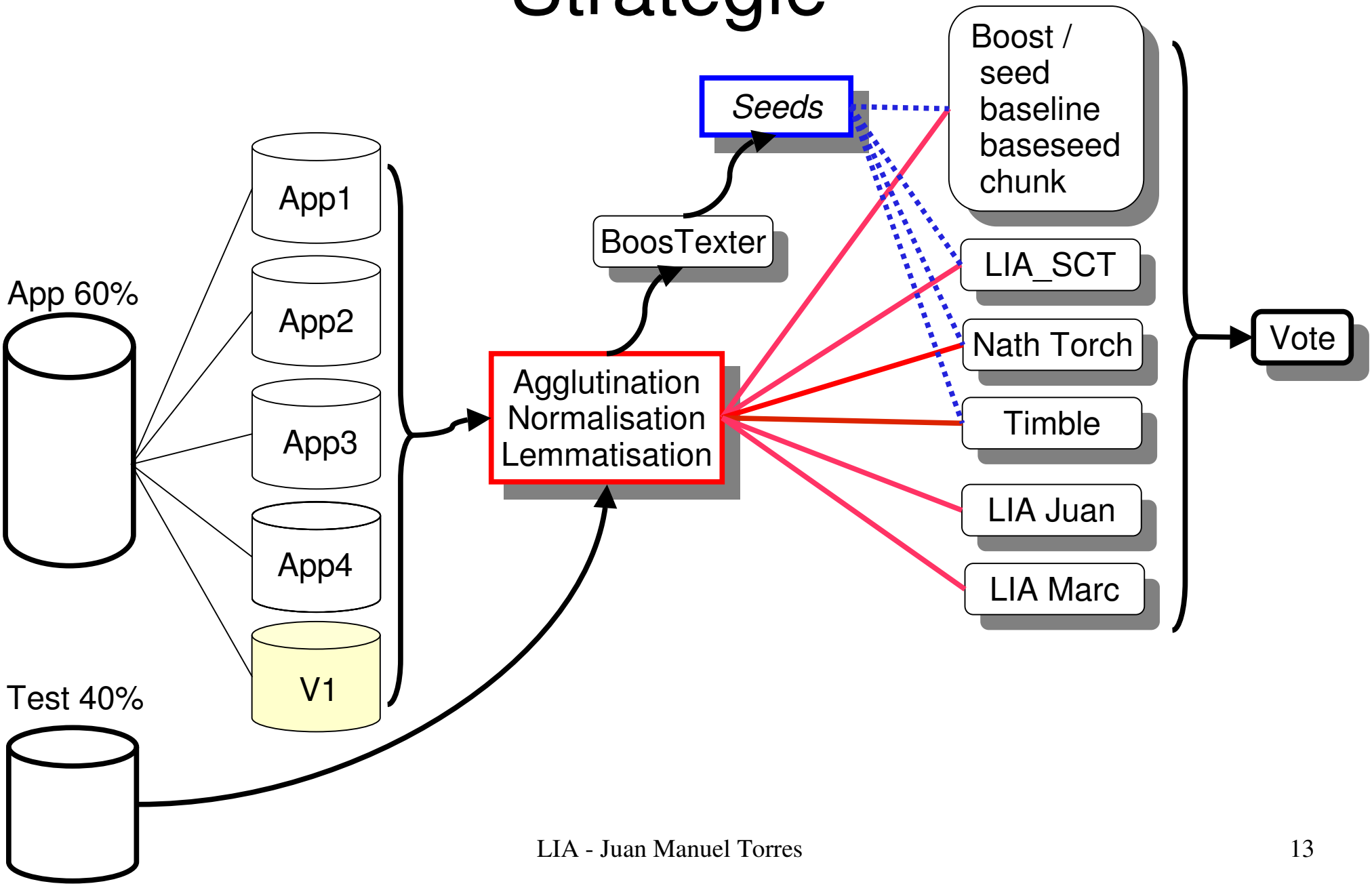
Déploiement des classifieurs (1/2)

- BoosTexter
  - Baseline : texte = sac de n-grammes sur les lemmes
  - Seed : texte = sac de n-grammes sur les seeds
  - BaseSeed : texte = sac de n-grammes sur les lemmes et les seeds
  - Chunk : texte = sac de n-grammes sur les lemmes limités aux «chunks» syntaxiques contenant un mot «seed»
- SVM-Torch : Texte = sac de seeds (unigrammes)
- Timble : Texte = sac de seeds (unigrammes)
- LIA_SCT : Suite de lemmes

Aucun « tuning » effectué pour : SVM-Torch, Timble et LIA_SCT !! (recette de base)

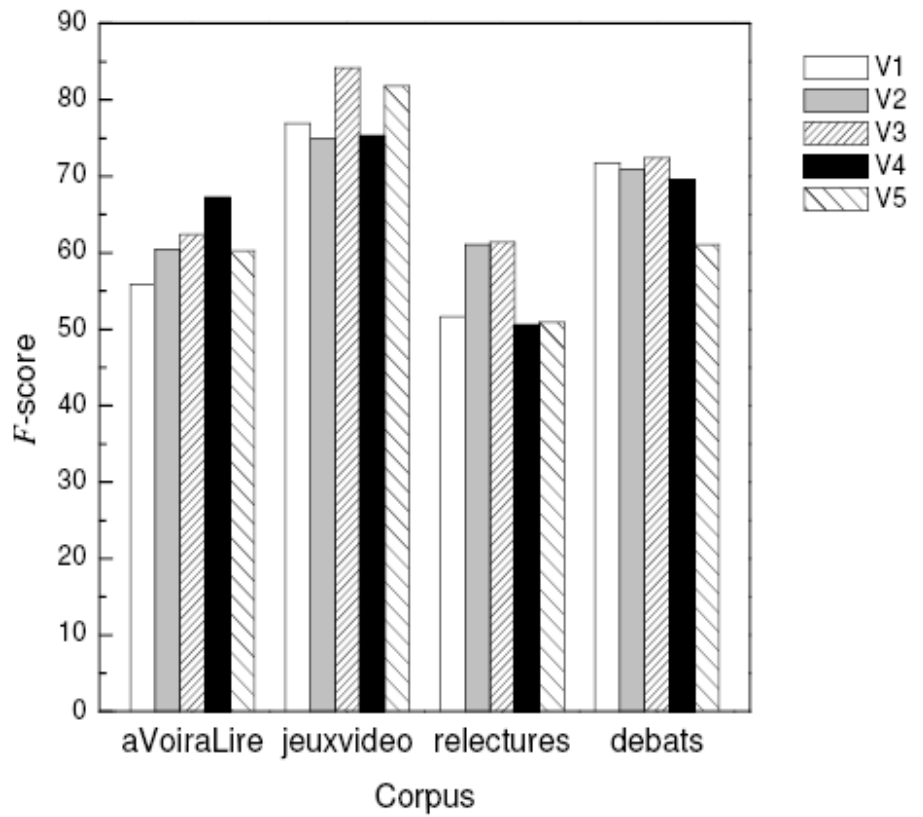
- LIA_JUAN : Suite de lemmes
- LIA_MARC : Suite de lemmes

Stratégie


App 60%
App1
App2
App3
App4
V1
Test 40%
Agglutination
Normalisation
Lemmatisation
BoosTexter
Seeds
Boost /
seed
baseline
baseseed
chunk
LIA_SCT
Nath Torch
Timble
LIA Juan
LIA Marc
Vote

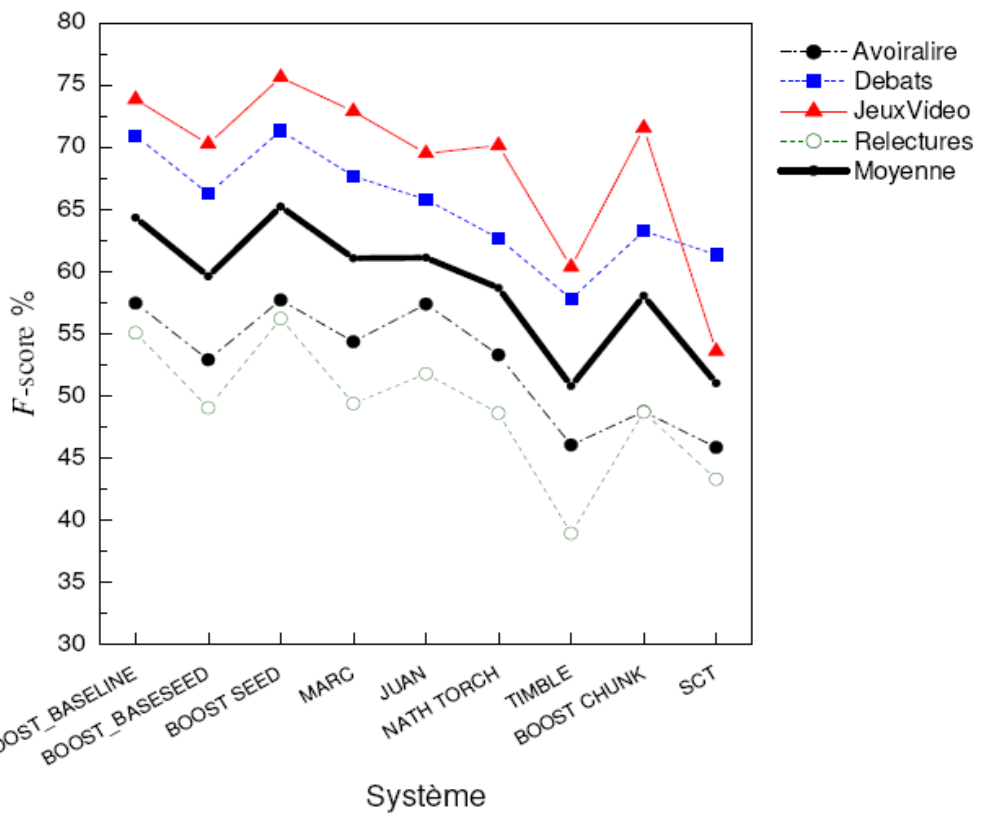
Corpus de validation

Corpus	Précision	Rappel	F-score	Correctes	Total
aVoiraLire (V)	0,6419	0,5678	0,6026	1 385	2 074
jeuxvideo (V)	0,8005	0,7730	0,7865	2 005	2 537
relectures (V)	0,5586	0,5452	0,5518	510	881
debats (V)	0,7265	0,7079	0,7171	12 761	17 299

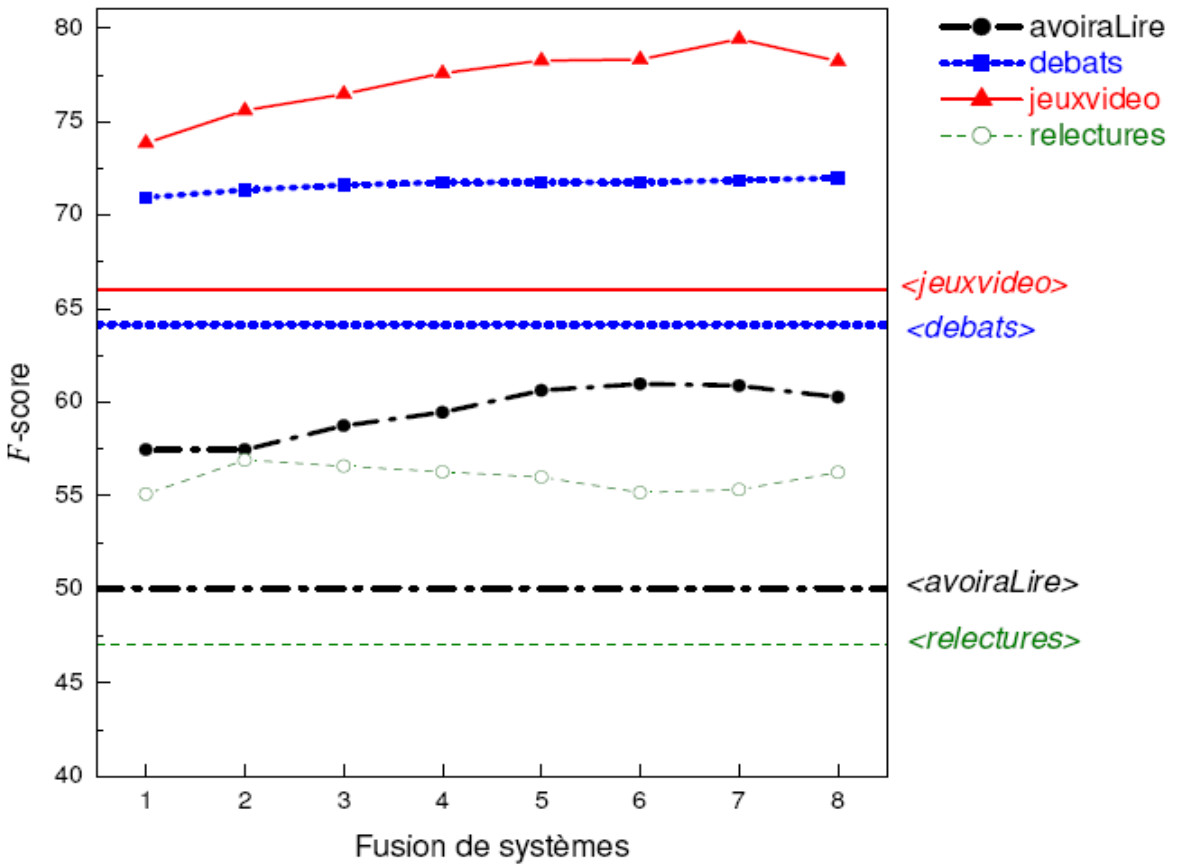


Corpus de test (1)

Corpus	Précision	Rappel	F-score	Correctes	Total
aVoiraLire (T)	0,6540	0,5590	0,6028	931	1 386
jeuxvideo (T)	0,8114	0,7555	0,7824	1 333	1 694
relectures (T)	0,5689	0,5565	0,5626	353	603
debats (T)	0,7307	0,7096	0,7200	8 403	11 533

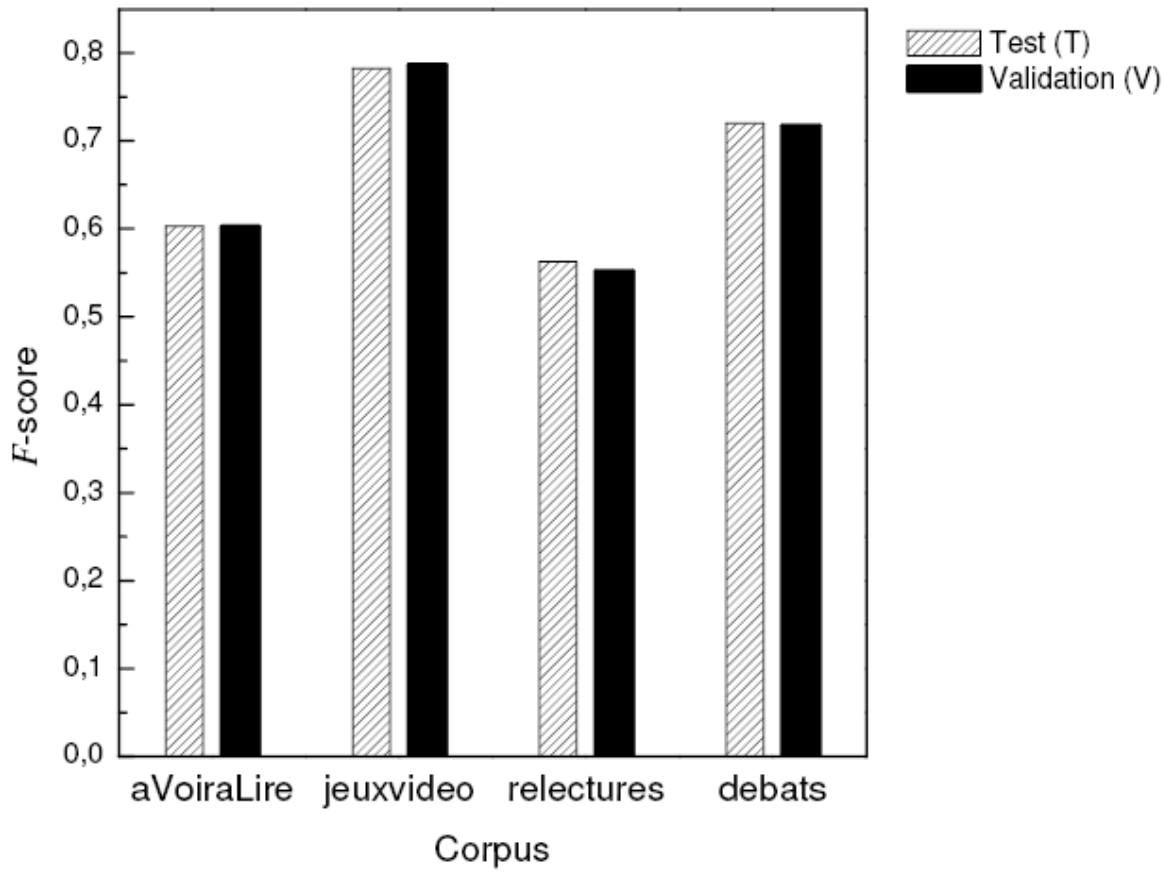


Corpus de test (2)



1 BOOST_BASELINE
2 BOOST_BASELINE+ BOOST_BASESEED + BOOST_SEED
3 BOOST_BASELINE+ BOOST_BASESEED + BOOST_SEED + MARC
4 BOOST_BASELINE+ BOOST_BASESEED + BOOST_SEED + JUAN
5 BOOST_BASELINE+ BOOST_BASESEED + BOOST_SEED + JUAN + NATH_TORCH
6 BOOST_BASELINE+ BOOST_BASESEED + BOOST_SEED + JUAN + NATH_TORCH + TIMBLE
7 BOOST_BASELINE+ BOOST_BASESEED + BOOST_SEED + JUAN + NATH_TORCH + TIMBLE + BOOST_CHUNK
8 BOOST_BASELINE+ BOOST_BASESEED + BOOST_SEED + JUAN + NATH_TORCH + TIMBLE + BOOST_CHUNK + SCT

Corpus de test (3)



Discussion

3 :2 relectures
L'idée d'appliquer les méthodes de classification pour définir des classes homogènes de pages web est assez originale par contre, la méthodologie appliquée est classique. Je recommande donc un « weak accept » pour cet article.

Le système l'a classé **1** (*accepté avec des modifications majeures*). De la lecture directe on pourrait en déduire que la classe est **1**... mais la référence est **2** (*accepté*)

3.6 relectures
Article trop court pour pouvoir être jugé. Je suggère de le mettre en POster si cela est prévu.

Document trop court : nos systèmes lui affectent l'etiquette **0** (rejet) mais il appartient à la classe **2** (acceptation pour les arbitres)

3 :9 relectures

Question : comment est construit le réseau bayesien ? Un peu bref ici... Remarques de forme : page 2, 4ème ligne, « comprend » 5ème ligne : "annotées" ou "annoté" page 3 : revoir la phrase confuse précédant le tableau dernière ligne, répétition de "permet" page 5 : 7ème ligne accorder "diagnostiqué" et "visé" avec "états" ou avec "connaissances"

Accepté... sur des remarques de forme uniquement !

aVoiraLire 1 :10

Depuis trente-six ans, chaque nouvelle production de David Bowie est un événement. Heathen , ne fait pas exception à cette règle. On reconnaît instantanément la patte de son vieux compère Tony Visconti. La voix de Bowie est mise en avant. Agréable surprise, surtout qu'elle n'a rien perdu depuis ses débuts. Là, commence le voyage. Ambiance, mélange dosé des instru[illegible]ès l'ouverture de l'album avec Sunday , un sentiment étrange nous envahit. Comme [illegible]Bowie venait de rentrer d'un voyage expérimental au coeur même de la musique[illegible]etour aux sources. L'ensemble du disque est rythmé par cette pulsation dont le d[illegible]a le secret. Le tout saupoudré de quelques pincées d'électronique. Le groupe est [illegible]duit au minimum. Outre Bowie en chef d'orchestre et Visconti, David Torn ponctue les compositions de ses guitares aventureuses et Matt Chamberlain apporte de l'âme [illegible]rythmique. Un quatuor à cordes fait une apparition, comme Pete Townshend (The [illegible]ho) ou Dave Grohl (ex-batteur de Nirvana). Avec trois reprises réarrangées et neuf compositions originales, le 25e album de Bowie est à l'image d'une cohérence artistique retrouvée.

Conclusions

- **L'union fait l'*intelligence* : combinaison de plusieurs méthodes numériques/probabilistes**
- **Problème trop simple ?**
- **Indépendance de la langue**
- **Indépendance du contexte**
- **Résultats bien au dessus de la moyenne**
- **Adaptable à d'autres problématiques**