

# Le langage naturel et la fouille de données

Mathieu Roche

Cours Fouille de Données  
Avancée

2007/2008

# Plan

- **Fouille de données et TAL**
  - Généralités
  - Les limites des approches actuelles
  - Approches pour améliorer les résultats
- **Evaluation des méthodes**

# Fouille de données et TAL

- **Types de méthodes :**
  - Statistiques
  - Linguistiques
  - Mixtes
- **Apprentissage supervisé et non supervisé**

# Fouille de données et TAL

- Limites liées aux langues étudiées
- Complexité du traitement du langage naturel (polysémie, traitement des anaphores, etc.)
- Quantité et qualité des données disponibles
- Qualité des systèmes de TAL

# Les améliorations possibles

- Ajouter des **connaissances sémantiques** (généralistes, spécialisées, comment obtenir ces informations ?)
- Ajouter des **connaissances lexicales**
- Ajouter des **connaissances syntaxiques**
- Leur combinaison ?

# Plan

- **Fouille de données et TAL**
  - Généralités
  - Les limites des approches actuelles
  - Approches pour améliorer les résultats
- **Evaluation des méthodes**

# Evaluation des méthodes

- Notion générale de **précision** et de **rappel**

$$précision = \frac{\text{nombre d'exemples positifs couverts}}{\text{nombre d'exemple couverts}}$$

**Une précision de 100% signifie que tous les exemples couverts sont positifs.**

$$rappel = \frac{\text{nombre d'exemples positifs couverts}}{\text{nombre d'exemples positifs}}$$

**Une couverture de 100% signifie que tous les exemples positifs sont couverts.**

# Evaluation des méthodes

- **Mesure pour combiner Rappel et Précision : la F-mesure (ou F-score).**

$$Fscore = \frac{(\beta^2 + 1) \times \text{Précision} \times \text{Rappel}}{\beta^2 \times \text{Précision} + \text{Rappel}}$$

→ Trois cas :  $\beta=1$ ,  $\beta<1$ ,  $\beta>1$

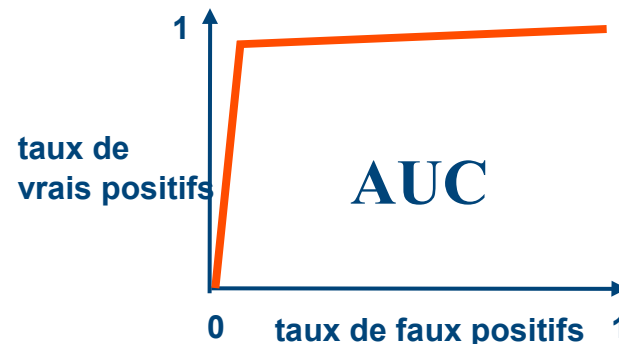
- **Variantes de la F-mesure**

# Evaluation des méthodes

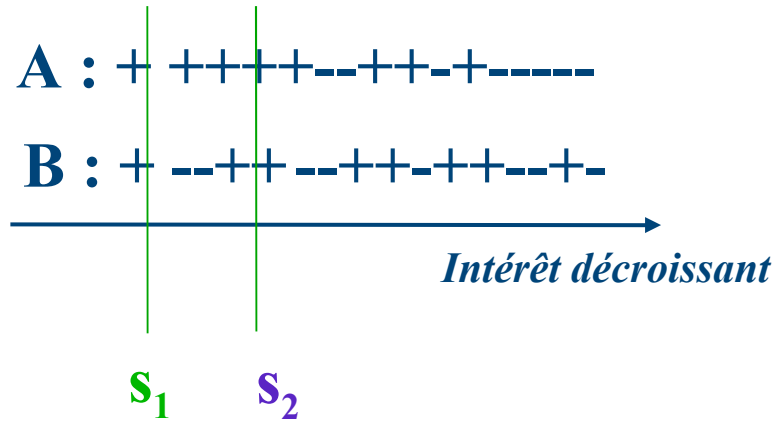
- **Front de Pareto (courbe Rappel/Précision) : plusieurs critères sont abordés**
- Le front de Pareto est défini par l'ensemble des approches telles qu'il n'existe pas une solution qui soit la meilleure pour tous les critères (ici précision et rappel).
- Les approches qui ne sont pas sur le front de Pareto sont dites “dominées”.
- Classement possible par front de Pareto

# Evaluation des méthodes

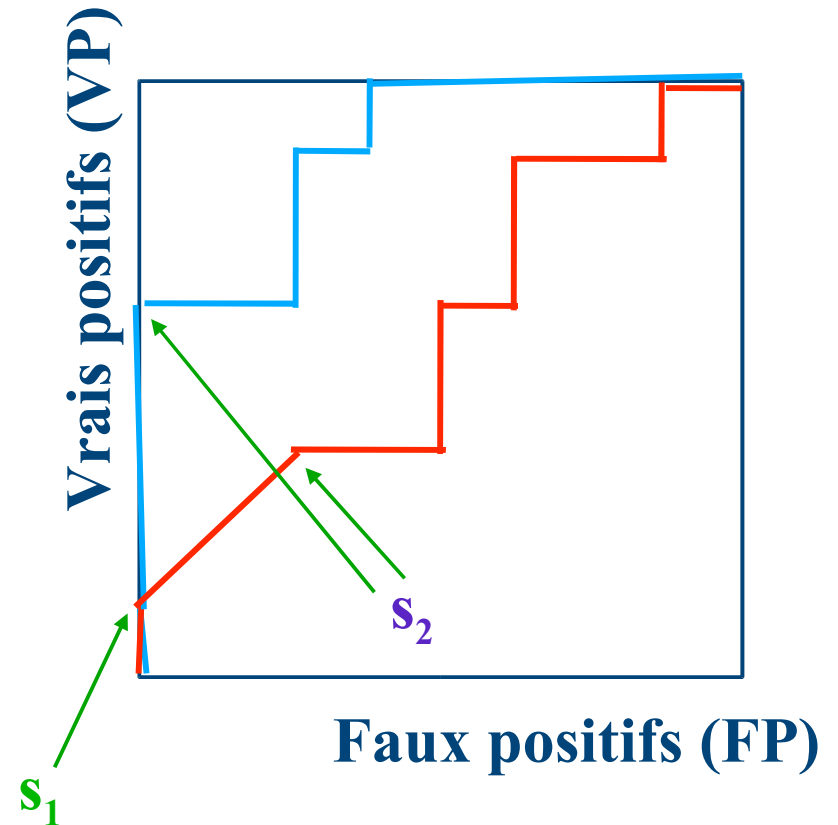
- Evaluation des fonctions de rang
- Utilisation des **courbes ROC** (Receiver Operating Characteristic) :  
courbe dont le taux de vrais positifs est représenté en ordonnées et  
le taux de faux positifs est représenté par l'axe des abscisses
- **Avantage** : **pas de sensibilité** dans le cas d'un **déséquilibre** entre les  
classes.



# Evaluation des méthodes



	$S_1$	$S_2$
A	VP = 1/8 FP = 0	VP = 5/8 FP = 0
B	VP = 1/8 FP = 0	VP = 3/8 FP = 2/8



# Evaluation de classifieurs (1/3)

## Évaluation du test : Matrice de confusion

		Réel	
		Pos	Neg
Prédit	Pos	TP	FP
	Neg	FN	TN

- TP : True Positive
- FP : False Positive
- FN : False Negative
- TN : True Negative

## Précision, Rappel, Accuracy

- $Precision = \frac{TP}{TP+FP}$
- $Rappel = \frac{TP}{TP+FN}$
- $Accuracy = \frac{TP+TN}{TP+FN+FP+TN}$

# Evaluation de classifieurs (2/3)

## Matrice de confusion multi-classes

		Réel					
		$C_1$	$C_2$	...	$C_i$	...	$C_n$
Prédit	$C_1$	$c_1^1$	$c_1^2$		$c_1^i$		$c_1^n$
	$C_2$	$c_2^1$					
	...			...			
	$C_i$	$c_i^1$			$c_i^i$		
	...			...			
	$C_n$	$c_n^1$					

- Prédiction correcte :  $c_i^i$
- Prédiction incorrecte :  $c_i^j$  avec  $i \neq j$

## Précision, Rappel, Accuracy

- $Precision(C_i) = \frac{c_i^i}{\sum_{j=1}^n c_i^j}$
- $Rappel(C_i) = \frac{c_i^i}{\sum_{j=1}^n c_j^i}$
- $Accuracy = \frac{\sum_{i=1}^n c_i^i}{\sum_{i,j=1}^n c_i^j}$

## Evaluation de classifieurs (3/3)

- La validation croisée sert à estimer l'erreur réelle d'un modèle (adaptée aux méthodes supervisées comme les KPPV)

Validation croisée (S, x)

// S est un ensemble x est un entier

Découper S en x parties égales  $\{S_1, \dots, S_x\}$

Pour i de 1 à x

Construire un modèle M avec l'ensemble S-S<sub>i</sub>

Evaluer l'erreur e<sub>i</sub> de M avec S<sub>i</sub>

Fin Pour

Retourner la moyenne des  $e_i = \frac{\sum_{i=1..x} e_i}{x}$

# Un exemple d'évaluation : le défi DEFT'06

Le thème général du défi DEFT'06 concernait la **reconnaissance automatique des segments thématiques** de textes écrits en français dans différents domaines.

La segmentation thématique peut être utilisée pour :

- isoler des zones répondant précisément à une requête (système de Q/R) ;
- l'indexation de textes ;
- classification de documents (prétraitement) ;
- le résumé de textes ;
- etc.

# Un exemple d'évaluation : le défi DEFT'06

## *3 corpus / 3 types de segments.*

**Corpus de discours politiques.** La segmentation thématique est fondée sur la structure thématique des discours mis en ligne sur le site de référence.

**Corpus de lois de l'Union Européenne.** Les segments thématiques sont les lois.

**Corpus scientifique.** Les segments thématiques à retrouver sont les différentes sections d'un ouvrage, à savoir les chapitres, sections, sous-sections et sous-sous-sections.

# Un exemple d'évaluation : le défi DEFT'06

**Corpus d'apprentissage** sont composés de **60% des corpus** associés. Ces corpus contiennent les informations permettant d'identifier les segmentations thématiques. Les participants ont eu trois mois pour mettre en place leurs méthodes de segmentation sur les corpus d'apprentissage.

Les **40% des corpus** restants ont été utilisés **pour le test**. Les participants ont eu deux jours pour appliquer, sur les corpus de test, les méthodes mises en oeuvre sur les corpus d'apprentissage.

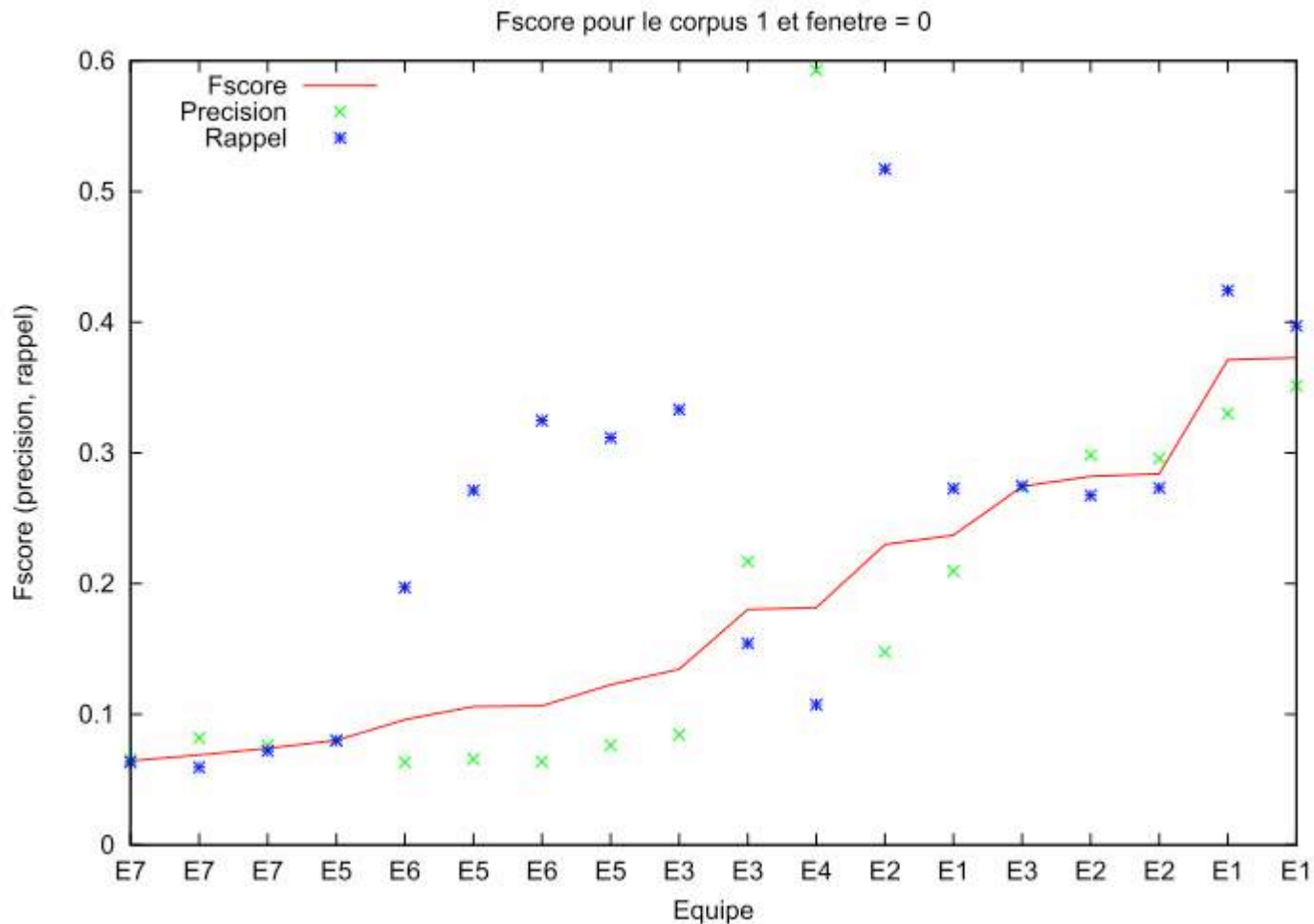
# Un exemple d'évaluation : le défi DEFT'06

## Evaluation :

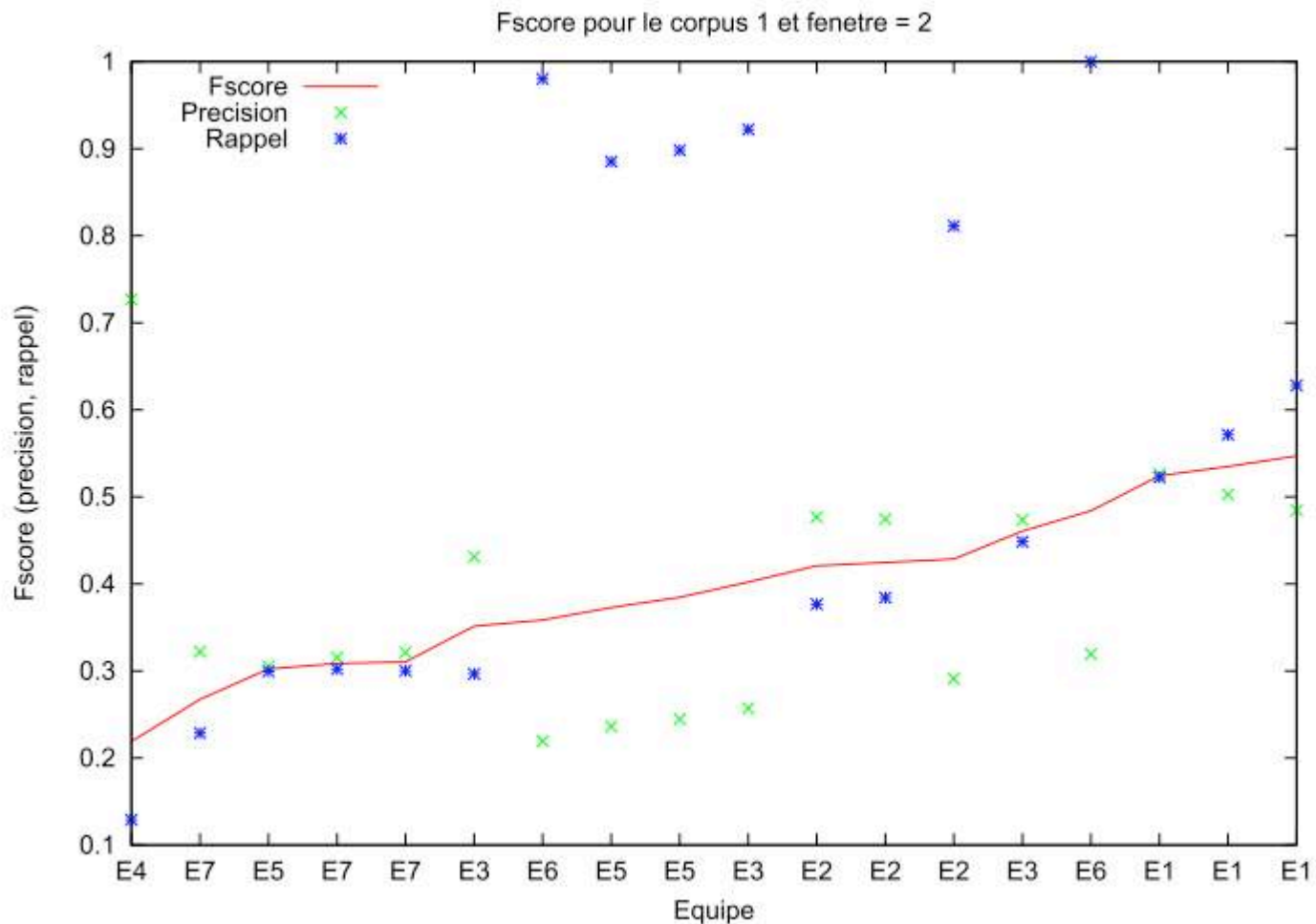
- **Précision, Rappel, F-mesure**

- **F-mesure floue** : marge d'erreur autour de la solution exacte. Une tolérance de  $\pm$  une (deux) phrase(s) autour de la phrase marquant le début du segment a été introduite.

# Un exemple d'évaluation : le défi DEFT'06 (Fmesure classique - corpus discours)



# Un exemple d'évaluation : le défi DEFT'06 (Fmesure floue - fenêtre 2 - corpus discours)



# Un exemple d'évaluation : le défi DEFT'06 (Front de Pareto - corpus discours)

