Résultats de l'édition 2007 du DÉfi Fouille de Textes

Les membres du Comité d'Organisation de DEFT'07 :
Jean-Baptiste Berthelin[1], Sarra El Ayari[1], Cyril Grouin[1], Thomas Heitz[2], Martine Hurault-Plantet[1], Michèle Jardino[1], Zohra Khalis[3] et Michel Lastes[1]

[1] LIMSI–CNRS,
{jean-baptiste.berthelin, sarra.elayari, cyril.grouin, martine.hurault-plantet, michel.lastes}@limsi.fr

[2] LRI, Université Paris-Sud,
heitz@lri.fr

[3] Épigénomique, Génopole d'Évry,
zkhalis@epigenomique.genopole.fr

Résumé : Cet article présente les résultats obtenus par chacun des participants à l'édition 2007 du Défi Fouille de Textes (DEFT). Ces résultats font apparaître une gradation des difficultés de traitement sur les différents corpus. Outre une vue d'ensemble des résultats, notre article décrit les méthodes retenues par les candidats lors des deux grands étapes du traitement : la *représentation* des textes et leur *classification*, pour laquelle les méthodes hybrides semblent prometteuses.

Mots-clés : F-score, rappel, précision, front de Pareto, tf*idf, représentation de textes, classification de textes.

Introduction

Pour cette édition du défi, chaque candidat avait la possibilité de soumettre jusqu'à trois résultats pour chacun des corpus. Chaque soumission a été considérée comme étant un ensemble indissociable portant sur les quatre corpus.

Pour toutes les soumissions, nous avons calculé le F-score strict (avec $\beta = 1$) puis, sur la base de ces calculs, nous avons défini la meilleure soumission de chaque équipe. Nous avons ensuite procédé au classement final des équipes en ne prenant en compte que la meilleure soumission de chacun des participants.

Les quatre corpus avaient été préalablement été soumis à des évaluations humaines, afin d'obtenir une approximation qualitative de la faisabilité d'une évaluation automatique. L'examen des résultats obtenus par les participants a été complété par celui des méthodes qu'ils ont employées, tant pour la représentation des textes que pour leur classification. Cette étude fait ressortir que la sélection des traits représentant chaque texte joue, dans ce cadre, un rôle crucial.

1 F-scores stricts

Au regard des résultats obtenus par chacun des participants sur chaque corpus (voir tableau n° 1), il apparaît assez nettement que les quatre corpus ont posé des problèmes distincts dans les traitements mis en œuvre. Nous pouvons ainsi établir un classement des corpus sur la base des F-scores obtenus, ces résultats traduisant les difficultés de traitement qu'ont rencontré les participants :

1. Corpus des jeux vidéos : les F-scores stricts des participants sont compris entre 0,784 et 0,457 ;
2. Corpus des débats parlementaires : les F-scores stricts sont compris entre 0,720 et 0,540 ;
3. Corpus « À voir, à lire » : les F-scores stricts sont compris entre 0,602 et 0,377 ;
4. Corpus des relectures : les F-scores stricts sont compris entre 0,566 et 0,398.

On observe les résultats les meilleurs pour les corpus des jeux vidéos et des débats parlementaires et, à l'inverse, de moins bons résultats pour les corpus des critiques et les relectures. Cette tendance semble partagée par l'ensemble des participants au défi comme l'attestent les graphiques n° 3 (F-scores stricts pour toutes les soumissions) et n° 4 (F-scores stricts pour les meilleures soumissions).

Équipe	Soumission	À voir, à lire	Jeux vidéos	Relectures	Débats
J.-M. Torres-Moreno (LIA)	1	0.602	0.784	0.564	0.719
J.-M. Torres-Moreno (LIA)	2	0.603	0.782	0.563	0.720
J.-M. Torres-Moreno (LIA)	3	0.603	0.743	0.566	0.709
G. Denhière (EPHE et U. Würzburg)	1	0.476	0.640	0.398	0.577
G. Denhière (EPHE et U. Würzburg)	2	0.599	0.699	0.507	0.681
S. Maurel (CELI France)	1	0.513	0.706	0.536	0.697
S. Maurel (CELI France)	2	0.418	0.538	0.477	0.697
S. Maurel (CELI France)	3	0.519	0.700	0.505	0.697
M. Vernier (GREYC)	1	0.577	0.761	0.414	0.673
M. Vernier (GREYC)	2	0.532	0.715	0.474	0.639
M. Vernier (GREYC)	3	0.532	0.715	0.474	0.673
E. Crestan (Yahoo ! Inc.)	1	0.529	0.670	0.441	0.652
E. Crestan (Yahoo ! Inc.)	2	0.523	0.673	0.462	0.703
M. Plantié (LGI2P et LIRMM)	1	0.421	0.783	0.478	0.618
M. Plantié (LGI2P et LIRMM)	2	0.424	0.732	0.442	0.671
M. Plantié (LGI2P et LIRMM)	3	0.472	0.547	0.442	0.608
A.-P. Trinh (LIP6)	1	0.542	0.659	0.427	0.676
A.-P. Trinh (LIP6)	2	0.490	0.580	0.467	0.665
M. Généreux (NLTG)	1	0.453	0.623	0.471	0.540
M. Généreux (NLTG)	2	0.464	0.626	0.463	0.554
M. Généreux (NLTG)	3	0.441	0.602	0.435	0.569
E. Charton (LIA)	1	0.377	0.619	0.433	0.616
E. Charton (LIA)	2	0.504	0.457	0.469	0.553
E. Charton (LIA)	3	0.504	0.619	0.419	0.553
A. Acosta (Lattice)	1	0.392	0.536	0.437	0.582

FIG. 1 – F-scores stricts ($\beta = 1$) pour toutes les soumissions sur chaque corpus.
La meilleure soumission de chaque équipe apparaît sur une ligne grisée.

Outre le fait que cette gradation de difficulté sur les différents corpus apparaît partagée par l'ensemble des participants au défi, nous remarquons également que ces résultats rejoignent les évaluations opérées par les juges humains (voir tableaux n° 2) :

1. Corpus des jeux vidéos : les F-scores stricts des juges humains sont compris entre 0,90 et 0,73 ;
2. Corpus « À voir, à lire » : les F-scores stricts sont compris entre 0,79 et 0,52 ;
3. Corpus des relectures : les F-scores stricts sont compris entre 0,58 et 0,41.

Les évaluateurs humains ont obtenu de meilleurs résultats sur les corpus des jeux vidéos et « *à voir, à lire* » que les systèmes automatiques des participants au défi. En revanche, les résultats sont quasi-identiques entre juges humains et systèmes automatiques sur le corpus des relectures, corpus jugé complexe par les humains.

Juge	1	2	3
F-score	0,73	0,86	0,90

Juge	1	2	3	4	5
F-score	0,52	0,76	0,69	0,70	0,79

Juge	1	2
F-score	0,41	0,58

FIG. 2 – F-scores obtenus par les juges humains sur les corpus « *à voir, à lire* » (tableau de gauche), des jeux vidéos (tableau central) et des relectures (tableau de droite).

Des méthodes d'analyse distinctes pour chaque type de corpus

Du fait de l'existence de corpus de différentes qualités littéraires (des phrases bien formulées dans les débats parlementaires aux phrases courtes et mal accentuées des relectures d'articles), des méthodes d'analyses distinctes ont été appliquées sur chaque corpus. Ces différences de méthodes ressortent dans les courbes des graphiques des F-scores stricts.

Si l'on considère qu'il existe une thématique « critiques » rassemblant les corpus des jeux vidéos et « *à voir, à lire* » (autrement dit, les critiques de livres et de films), il semblerait – d'après le graphique n° 3 – que les candidats

ont chacun appliqué la même méthode sur ces deux corpus ; pour une soumission donnée, on trouvera ainsi le même type de résultats (bon ou mauvais) pour ces deux corpus. Il en résulte que ces deux courbes évoluent globalement en parallèle sur les deux graphiques.

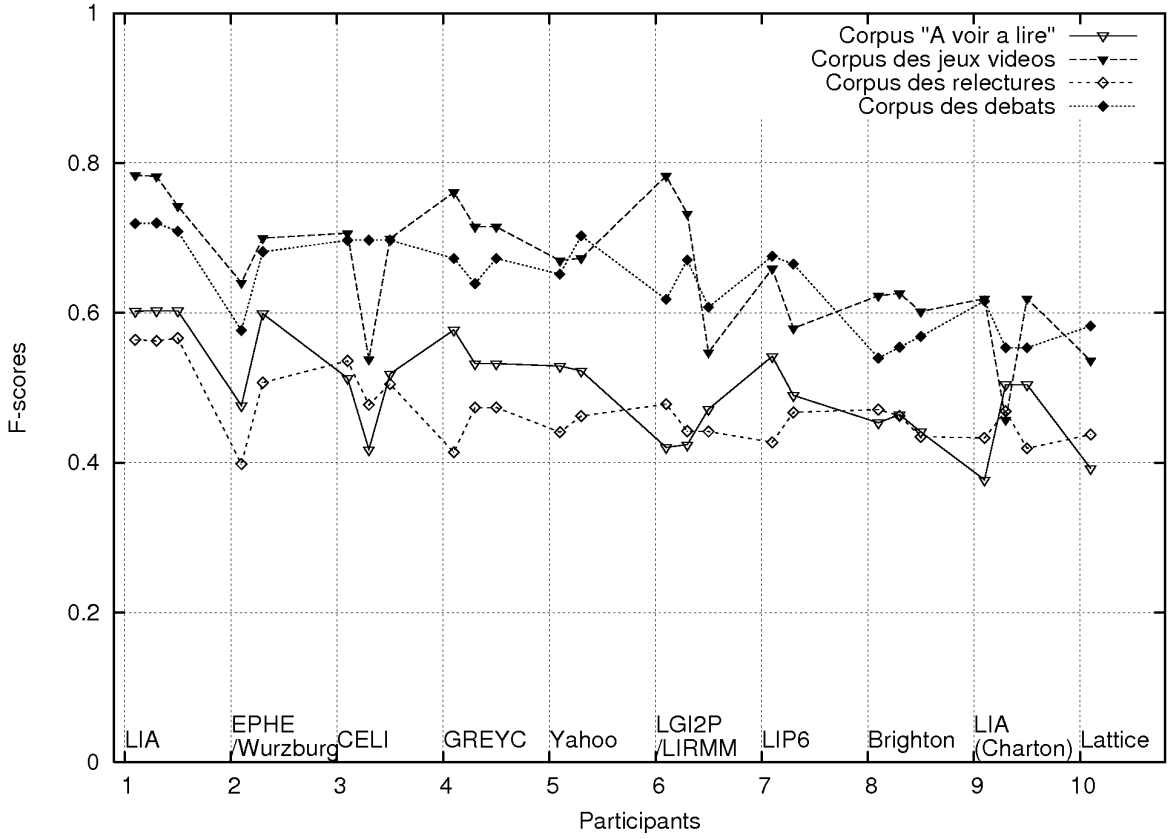

FIG. 3 – F-score strict ($\beta = 1$) pour l'ensemble des soumissions de chacun des candidats.

Un corpus difficile : les relectures d'articles

Malgré les difficultés rencontrées sur le corpus des relectures, quelques équipes semblent avoir eu moins de difficultés pour ce corpus que pour celui des critiques de livres et de films. Il en est ainsi pour deux soumissions du CELI, mais également de l'équipe LGI2P/LIRMM, de l'équipe de Michel Généreux (noté « Brighton »), et de deux des trois équipes jeunes chercheurs : une soumission pour Eric Charton et la soumission du Lattice.

Un corpus apprécié des équipes jeunes chercheurs : les débats parlementaires

Si l'on considère les équipes jeunes chercheurs indépendamment des autres équipes, une singularité émerge quant au corpus des débats parlementaires. Alors que les meilleurs résultats ont été obtenus sur le corpus des jeux vidéos, les équipes de jeunes chercheurs (notées « LIP6 », « LIA (Charton) » et « Lattice » en légende des graphiques) ont obtenu leurs meilleurs résultats sur le corpus des débats parlementaires.

Pour le cas où ces équipes auraient soumis plusieurs résultats, la meilleure soumission de chacune de ces équipes demeure celle où les résultats sur le corpus des débats parlementaires sont les meilleurs. Cette constatation s'avère assez flagrante sur le graphique n° 4. À ce titre, l'équipe « Yahoo » est la seule équipe hors catégorie « jeunes chercheurs » à avoir pour meilleure soumission celle où les résultats obtenus sur le corpus des débats parlementaires sont les plus élevés.

L'incidence de l'indice de confiance sur les résultats

Les participants ont eu la possibilité d'associer un indice de confiance à chaque note attribuée aux documents des corpus. Cet indice de confiance était proposé de manière optionnelle.

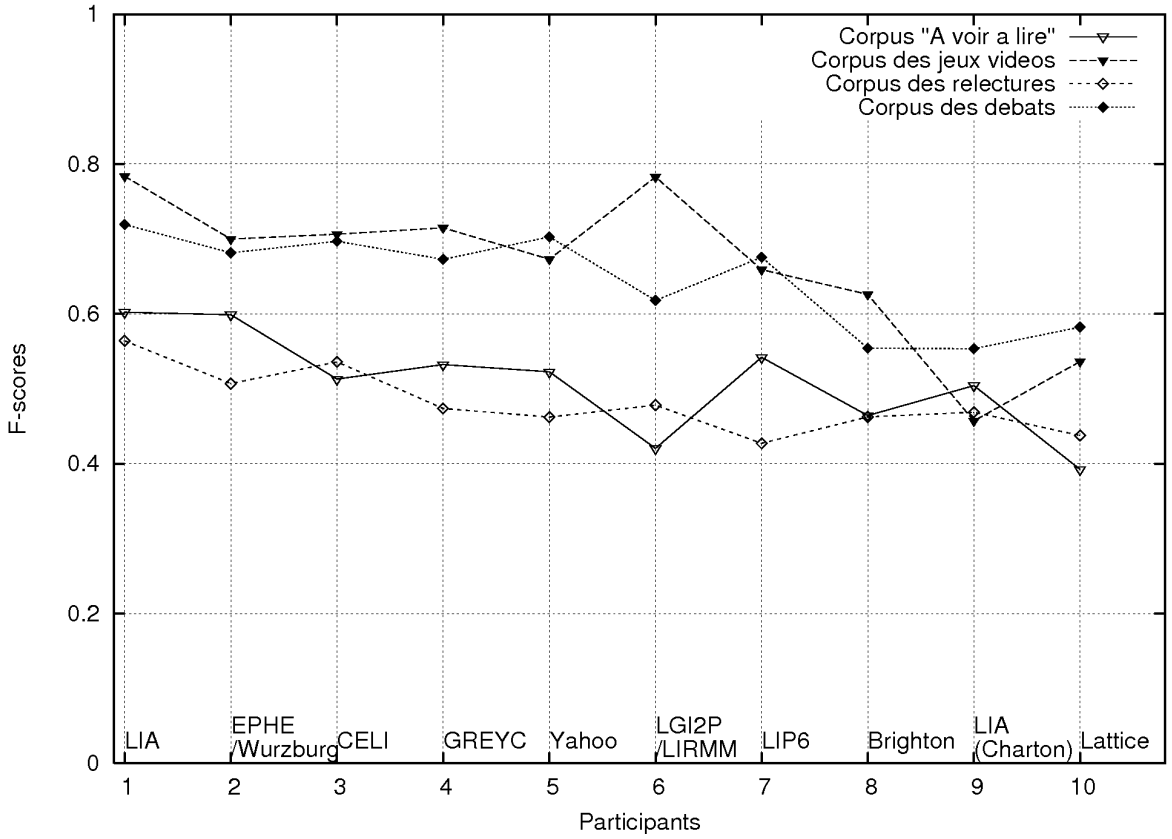

FIG. 4 – F-score strict ($\beta = 1$) pour les meilleures soumissions de chacun des candidats.

Sur les dix participants au défi, six y ont recouru. Sur ces six participants, certains l'ont appliquée pour chaque soumission, d'autres n'ont proposé que certaines soumissions avec indice de confiance (en générale deux soumissions avec indice de confiance, une soumission sans) :

- **Soumissions avec indice de confiance :** LIA (soumissions n° 1, 2 et 3), CELI France (n° 1, 2 et 3), GREYC (n° 1 uniquement), LGI2P/LIRMM (n° 1 et 3), LIP6 (n° 1 et 2) et M. Généreux (n° 1, 2 et 3) ;
- **Soumissions sans indice de confiance :** LPC/UP8 (soumissions n° 1 et 2), GREYC (n° 2 et 3), Yahoo ! Inc. (n° 1 et 2), LGI2P/LIRMM (n° 2 uniquement), E. Charton (n° 1, 2 et 3) et Lattice (n° 1).

Les résultats ne nous permettent pas d'établir une corrélation entre les scores obtenus et l'utilisation de l'indice de confiance dans les notes attribuées.

2 Front de Pareto

Définition

Le front de Pareto est défini par l'ensemble des approches qui sont telles qu'aucune autre approche ne présente de meilleurs résultats pour tous les critères étudiés, en l'occurrence le rappel et la précision.

Représentation graphique

Le rappel est présenté sur l'axe des abscisses, la précision sur l'axe des ordonnées. Les courbes correspondent aux valeurs de F-score comprises entre 0,1 et 0,9 (avec $\beta = 1$).

Le front de Pareto est symbolisé sur ces schémas par l'ensemble des points qui sont reliés par des tirets. Les points isolés sont donc exclus du front de Pareto.

Les numéros aux côtés des points permettent d'identifier les équipes, un point représentant une soumission pour le corpus considéré (notez que le numéro de la soumission n'apparaît pas sur ces schémas) :

Numéro	Équipe
3	M. Généreux (NLTG–Université de Brighton)
4	M. Plantié (LGI2P et LIRMM)
5	G. Denhière (LPC–Université de Provence et Université Paris 8)
6	M. Vernier (GREYC)
7	E. Crestan (Yahoo ! Inc.)
8	A.-P ; Trinh (LIP6), *équipe jeunes chercheurs*
9	A. Acosta (Lattice), *équipe jeunes chercheurs*
11	J.-M. Torres-Moreno (LIA)
13	S. Maurel (CELI France)
14	E. Charton (LIA), *équipe jeunes chercheurs*

Une difficulté partagée sur certains corpus

L'analyse des histogrammes n° 5 à 8 met en avant plusieurs éléments. En premier lieu, la difficulté partagée par l'ensemble des participants sur certains corpus, en particulier celui des relectures (figure n° 8) pour lequel les résultats sont moins bons que pour les autres corpus, avec des valeurs de rappel, de précision et de F-score strict qui s'échelonnent entre 0,4 et 0,6 sans trop de disparités entre chaque candidat.

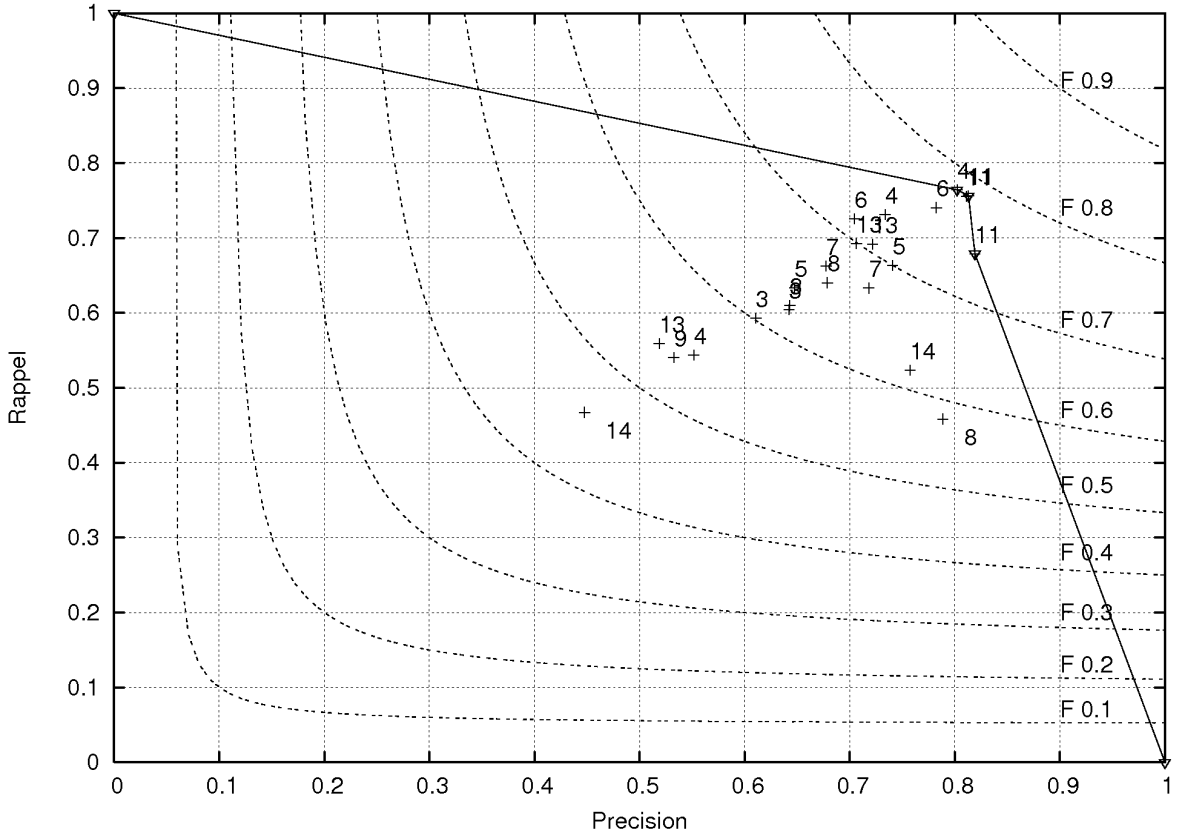

FIG. 5 – Front de Pareto pour le corpus des jeux vidéos.

A contrario, l'histogramme n° 5 confirme la bonne réussite des méthodes d'analyse du corpus des jeux vidéos, corpus pour lequel la majorité des valeurs de rappel, précision et F-score strict dépasse 0,5. La réussite sur le corpus des débats parlementaires est également visible sur l'histogramme n° 6 où les valeurs de rappel, de précision et de F-score strict sont comprises entre 0,5 et 0,75.

Homogénéité et hétérogénéité des résultats

Il est possible de tirer un second enseignement de la part de ces histogrammes : celui de l'homogénéité des résultats entre participants.

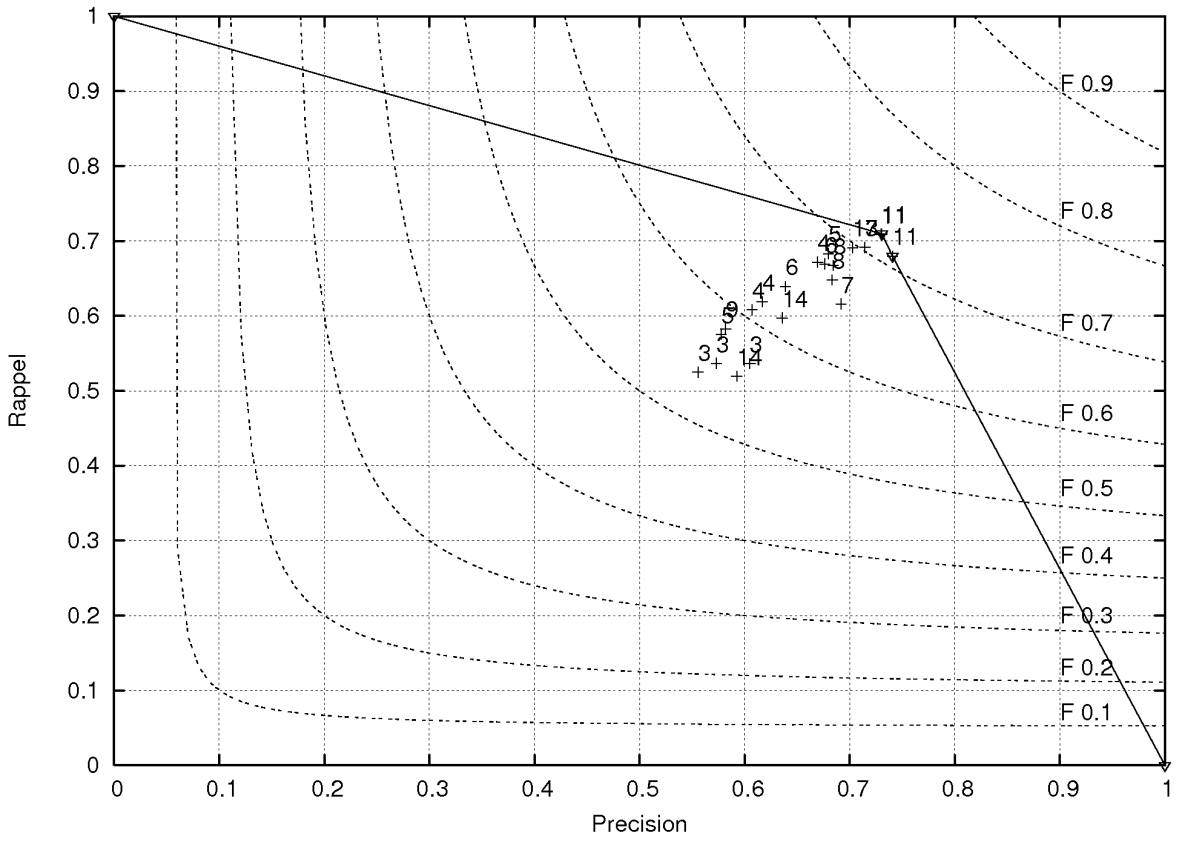

FIG. 6 – Front de Pareto pour le corpus des débats parlementaires.

Alors que les résultats sur les corpus des jeux vidéos et « *à voir, à lire* » sont assez hétérogènes selon les candidats et les soumissions et apparaissent clairsemés dans ces histogrammes (figures n° 5 et 7), les résultats portant sur les corpus des débats parlementaires et des relectures sont davantage homogènes et se présentent sous la forme de nuages de points assez compacts sur les histogrammes (figures n° 6 et 8).

Deux interprétations sont possibles pour ces nuages de points : soit le corpus était difficile à analyser et les résultats sont tous moyens (c'est semble-t-il le cas pour le corpus des relectures), soit au contraire le corpus était facile à analyser et les résultats ne pouvaient dès lors qu'être bons (voir tableau n° 1) ; pour ce dernier cas, c'est – nous l'espérons et le supposons – le cas du corpus des débats parlementaires où deux arguments militent en faveur de cette analyse : d'une part, l'échelle de notes réduites à deux classes (pour ou contre) et d'autre part, la qualité littéraire des retranscriptions des débats (assez peu de fautes d'orthographe et de grammaire en comparaison du corpus des relectures) permettant d'appliquer des traitements robustes.

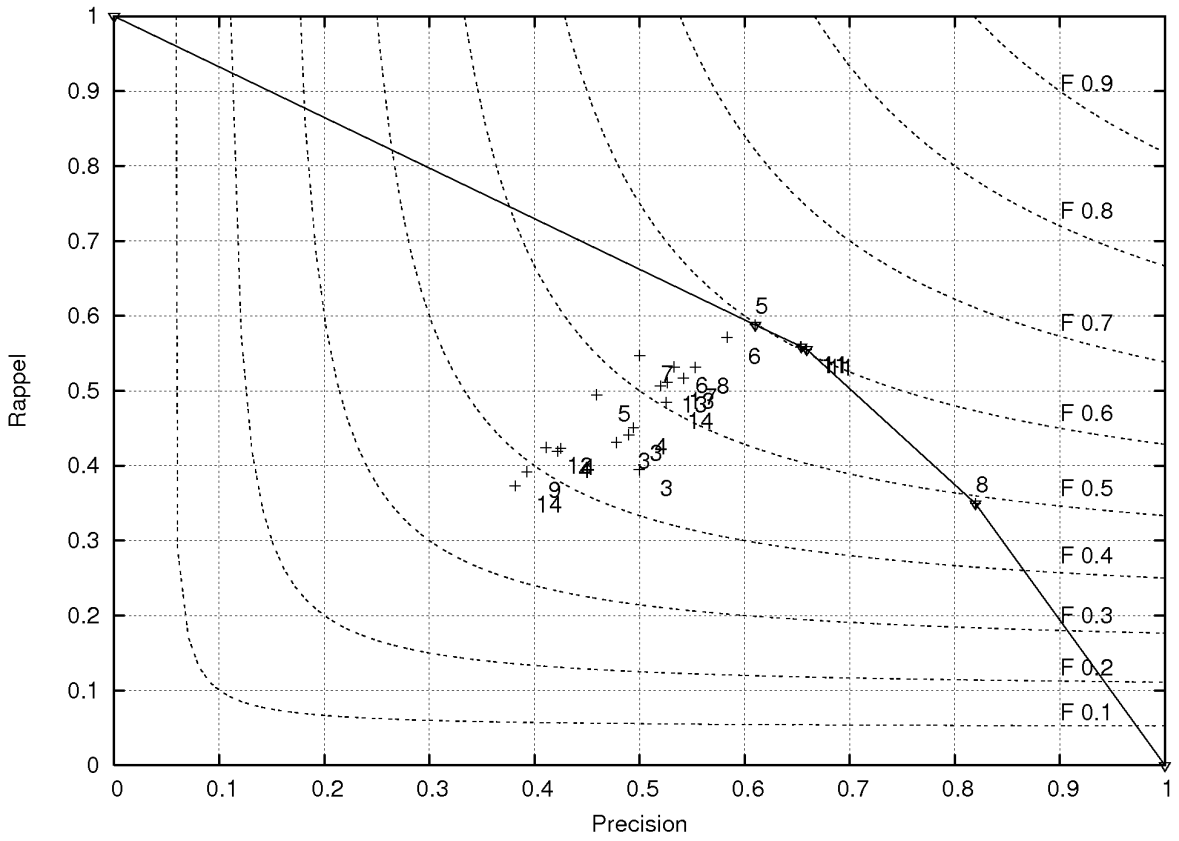

FIG. 7 – Front de Pareto pour le corpus « *à voir, à lire* ».

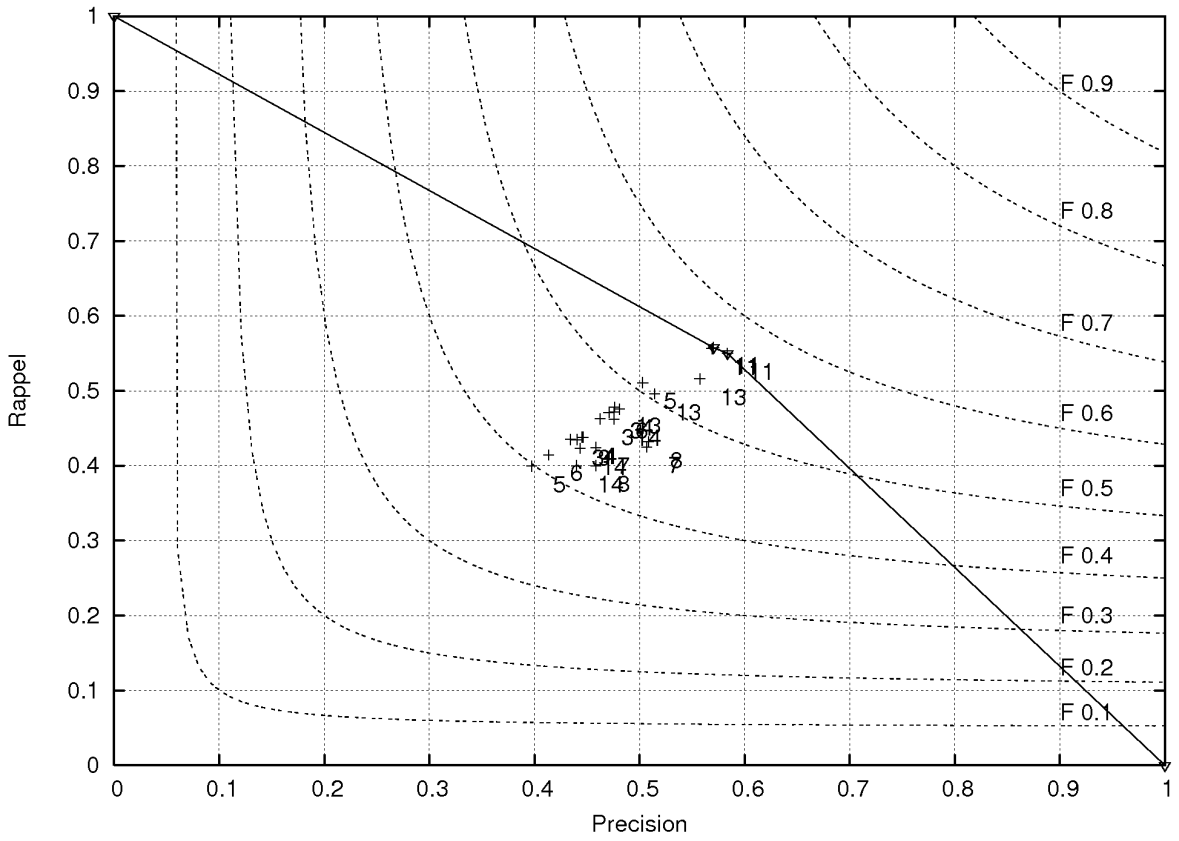

FIG. 8 – Front de Pareto pour le corpus des relectures.

3 Les méthodes utilisées par les participants

On peut séparer le processus généralement utilisé en deux grandes étapes : une première étape de représentation du texte, et une deuxième étape de classification.

L'étape de représentation du texte peut être plus ou moins élaborée, mais elle aboutit toujours à une réduction, parfois drastique, de l'ensemble des traits pouvant représenter les textes. Certaines équipes ont choisi de ne retenir qu'une partie du texte : les segments qui leur paraissaient pertinents pour l'évaluation de l'opinion. Ce peut être des paragraphes délimités tels que l'introduction et la conclusion du texte (Lattice, LIA-jeunes), ou bien des extraits trouvés par une méthode d'extraction de relations d'opinion (CELI-France), ou encore l'extraction du segment de texte autour d'un mot attracteur (LIA-jeunes). Dans le même esprit, l'équipe du GREYC-CRISCO a choisi de donner des poids différents aux différentes parties du texte.

Les méthodes de sélection des traits représentant les textes ont également été très variées. Plusieurs équipes ont utilisé un vocabulaire d'opinion (LIA, CELI-France, NLTG-Brighton, GREYC-CRISCO) soit pour pondérer les termes d'opinion dans les textes, soit pour les sélectionner. Les termes du domaine du corpus (par exemple article ou film) ont également été utilisés comme des attracteurs ou sélecteurs de termes d'opinion pertinents (LIA-jeunes, CELI-France, GREYC-CRISCO). L'équipe du Lattice a ajouté comme traits des statistiques sur les parties du discours. L'équipe du EPHE-CNRS et Universität-Würzburg a construit des concepts par analyse sémantique latente. Enfin, plusieurs participants ont utilisé plus classiquement une discrimination des traits importants pour chaque classe ou chaque texte par un critère statistique ou probabiliste tel que tf*idf, gain d'information, ou information mutuelle (LIA-jeunes, Lattice, Yahoo !Inc., LIP6, NLTG-Brighton).

L'étape de classification est également riche en méthodes différentes. Le classifieur le plus utilisé a été la machine à vecteur de support (SVM), mais ce n'est pas celui qui a produit les meilleurs résultats. Certaines équipes ont conçu des méthodes hybrides utilisant au moins deux classifieurs (LIA, LGI2P-LIRMM, CELI-France). C'est l'équipe du LIA qui a poussé le plus loin la méthode en prenant 6 classifieurs avec des variantes dans la représentation du texte donnant 9 systèmes de décision, un même poids étant attribué à chaque système dans la fusion finale. Cette méthode a produit les meilleurs scores. Une autre méthode utilisée plusieurs fois avec un certain succès a été la sommation de scores calculés sur chaque terme d'un document (Yahoo !Inc., LIA-jeunes, GREYC-

CRISCO) ou sur chaque relation d'opinion (CELI-France). Parmi les autres méthodes de classification on trouve les arbres de décision (Lattice, LIA, LGI2P-LIRMM), la régression logistique (LIA-jeunes, Lattice), des méthodes probabilistes (LIA, LGI2P-LIRMM, CELI-France), des réseaux de neurones (LGI2P-LIRMM), un algorithme de boosting (LIA), l'algorithme des k plus proches voisins (LIA), un classifieur à base de règles d'association (GREYC-CRISCO), et un calcul de similarité entre le vecteur représentant une classe et le vecteur représentant un texte (LIA-jeunes, EPHE-CNRS et Universität-Würzburg).

Conclusion

Les résultats ont été bons dans l'ensemble et finalement assez proches les uns des autres. Les résultats des tests faits avec les juges humains sont légèrement supérieurs mais montrent le même ordre de difficulté dans le traitement des corpus que les méthodes automatiques : le corpus des relectures est le plus difficile à évaluer, et celui des tests des jeux vidéos le plus facile. Les participants ont utilisé des méthodes très variées allant des approches statistiques à des approches linguistiques, syntaxiques ou sémantiques. L'utilisation d'un vocabulaire d'opinion a produit de bons résultats. La sélection des traits représentant le texte semble, au vu des résultats, presque plus importante que la classification proprement dite. Par ailleurs, les méthodes hybrides de classification semblent prometteuses.