Résultats de l'édition 2008 du DÉfi Fouille de Textes

Martine Hurault-Plantet[1] Jean-Baptiste Berthelin[1] Sarra El Ayari[1]
Cyril Grouin[1] Sylvain Loiseau[1] Patrick Paroubek[1]
(1) LIMSI-CNRS – BP133 – 91403 Orsay Cedex
{martine.hurault-plantet, jean-baptiste.berthelin, sarra.elayari,
cyril.grouin, sylvain.loiseau, patrick.paroubek}@limsi.fr

Résumé. Cet article présente les résultats obtenus par les participants de l'édition 2008 du défi fouille de textes (DEFT). Ces résultats se révèlent particulièrement élevés et homogènes entre chaque participant, avec une réussite accrue sur l'identification du genre par opposition à l'identification des thèmes. Dans cet article, nous revenons sur l'ensemble des résultats en opposant les F-scores stricts aux F-scores de confiance ; nous mettons également en avant l'incidence du score de confiance sur les résultats. Enfin, nous présentons les méthodes utilisées par les participants.

Abstract. This article presents the results obtained by the participants to the 2008 edition of the DEFT text-mining challenge. These results appear to be both high and homogeneous between any two participants, and successes are greater in genre identification as aopposed to topic identification. In this article, we survey the totality of the results, contrasting strict F-scores with confidence F-scores ; we also emphasize the incidence of confidence F-scores upon results. Finally, we present the methods used by the participants.

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

Keywords: F-score, recall, precision, Pareto front, tf*idf, text representation, text classification.

Introduction

Comme lors de la précédente édition du défi, chaque candidat avait la possibilité de soumettre jusqu'à trois résultats. Chaque soumission a été considérée comme étant un ensemble indissociable portant sur les deux tâches. Chaque soumission comportait donc 3 résultats de catégorisation : la catégorie en genre et la catégorie thématique des documents du corpus de la tâche 1 et la catégorisation thématique des documents du corpus de la tâche 2.

Pour toutes les soumissions, nous avons calculé le F-score strict (avec $\beta = 1$) pour chaque résultat de catégorisation 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.

1 F-scores stricts

1.1 Résultats des participants

Cette nouvelle édition du défi a révélé l'excellence des résultats obtenus par l'ensemble des participants – hormis quelques rares accidents – et ce, quelle que soit la sous-tâche considérée (tableau n° 1).

Équipe par ordre d'inscription	Soumission	T1 genre	T1 cat	T2 cat	Confiance
J. M. Torres-Moreno (LIA)	1	0.958	0.859	0.859	oui
J. M. Torres-Moreno (LIA)	2	0.981	0.883	0.872	oui
J. M. Torres-Moreno (LIA)	3	0.980	0.854	0.880	oui
M. Plantié (LGI2P/LIRMM)	1	0.971	0.853	0.858	non
M. Plantié (LGI2P/LIRMM)	2	0.970	0.852	0.852	non
M. Plantié (LGI2P/LIRMM)	3	0.955	0.823	0.828	non
E. Charton (LIA)	1	0.980	0.875	0.879	non
E. Charton (LIA)	2	0.959	0.809	0.662	non
E. Charton (LIA)	3	0.980	0.844	0.853	non
D. Buffoni (LIP6)	1	0.951	0.804	0.874	oui
D. Buffoni (LIP6)	2	0.973	0.879	0.874	oui
D. Buffoni (LIP6)	3	0.976	0.894	0.876	oui
G. Cleuziou (LIFO/INaLCO)	1	0.937	0.790	0.821	non
F. Rioult (GREYC)	1	0.964	0.849	0.838	oui
F. Rioult (GREYC)	2	0.856	0.672	0.328	non
F. Rioult (GREYC)	3	0.964	0.672	0.815	non

FIG. 1 – F-scores stricts ($\beta = 1$) pour toutes les soumissions sur chaque tâche avec indication d'utilisation de l'indice de confiance dans les résultats soumis. La meilleure soumission de chaque équipe apparaît sur une ligne grisée.

Les résultats sur l'identification du genre, certes limitée à deux choix possibles, se sont révélés excellents. L'identification des catégories a également produit de très bons résultats, comme l'attestent les F-scores stricts obtenus par les participants :

1. Identification du genre (tâche n° 1) : les F-scores stricts des participants sont compris entre 0,856 et 0,981 avec trois soumissions où le F-score strict s'établit à 0,980 ;
2. Identification de la catégorie (tâche n° 1) : les F-scores sont compris entre 0,672 et 0,894 ;
3. Identification de la catégorie (tâche n° 2) : les F-scores sont compris entre 0,328 et 0,880.

Pour chaque tâche, nous constatons l'extrême homogénéité des résultats entre les différents participants. En dehors de quelques rares soumissions moins bonnes, les F-scores stricts se tiennent dans des intervalles assez restreints, comme en témoignent les nuages de points représentés sur les graphiques du Front de Pareto (voir graphiques n° 5, 6 et 7).

En comparaison avec l'édition 2007 de DEFT (Paroubek *et al.*, 2007), il apparaît que les résultats des participants sont meilleurs cette année, et surpassent même les évaluations manuelles effectuées par les organisateurs.

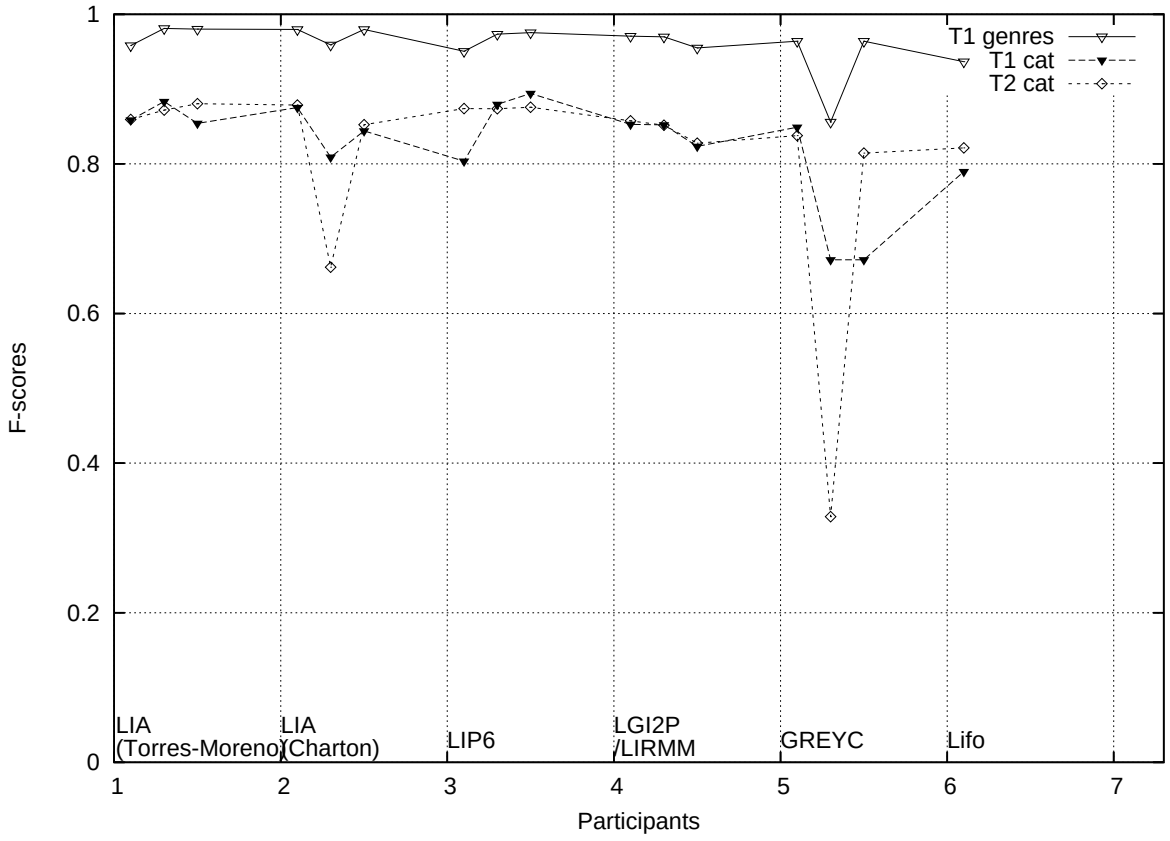

FIG. 2 – F-score strict ($\beta = 1$) pour l'ensemble des soumissions de chacun des candidats.

1.2 De meilleurs résultats que les juges humains...

Lors de la préparation de la tâche, nous avons procédé à des tests d'évaluation entre juges humains. Ces tests se sont révélés particulièrement bons et nous ont confortés dans le choix de cette tâche. À la réception des résultats des participants, force est de constater la supériorité des machines au regard des résultats obtenus par ces dernières sur les évaluateurs humains ! Cette constatation se révèle davantage sur l'identification des catégories.

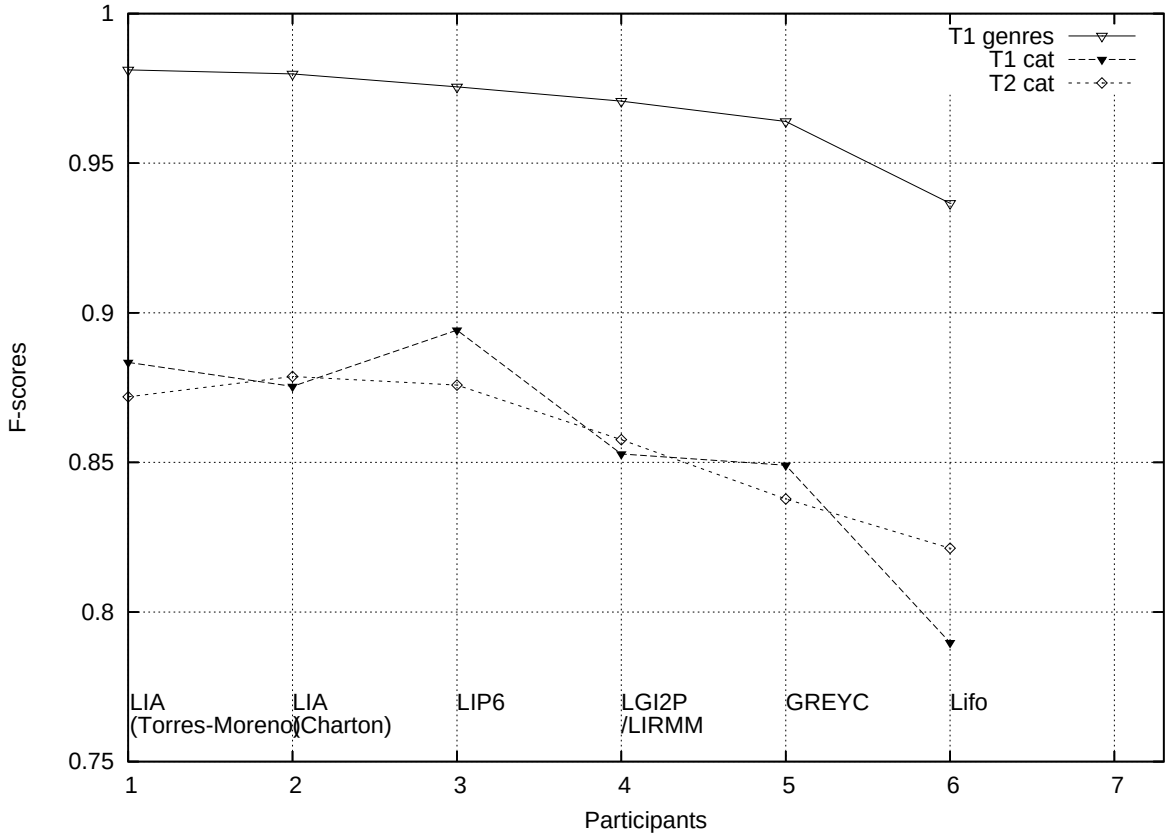

FIG. 3 – F-score strict ($\beta = 1$) pour les meilleures soumissions de chacun des candidats.

Au niveau de l'identification du genre, les F-scores stricts évoluent entre 0,94 et 1,00 pour les juges humains et entre 0,95 et 0,98 pour les meilleures soumissions des participants au défi (voir tableau 1). Concernant l'identification des catégories, les F-scores stricts s'échelonnent entre 0,66 et 0,82 pour les évaluateurs humains contre 0,84 à 0,89 pour les meilleures soumissions (fusion des résultats obtenus pour l'identification des catégories des tâches 1 et 2).

Par cette comparaison – au-delà du nombre réduit de choix possibles dans chacune des sous-tâches –, l'intérêt d'utiliser l'outil informatique dans le domaine de la classification thématique et de l'identification de contextes d'écritures différents apparaît de manière assez évidente.

2 F-score de confiance

2.1 Résultats

L'utilisation de l'indice de confiance dans les résultats soumis par les participants pour les catégories thématiques modifie les valeurs du F-score de manière différente selon les équipes. On observe ainsi une gradation de la baisse de la valeur du F-score pondéré par rapport au F-score strict, avec une répartition assez nette des soumissions entre celles dont le F-score n'a pas ou peu baissé et celles pour lesquelles le F-score a fortement baissé :

– pas de baisse ou faible baisse : 3 soumissions sont concernées ;

– baisse moyenne : une soumission concernée (première soumission de J. M. Torres-Moreno) ;
– forte baisse : 3 soumissions concernées.

Équipe par ordre d'inscription	Soumission	T1 cat	T2 cat
J. M. Torres-Moreno (LIA)	1	0.639	0.633
J. M. Torres-Moreno (LIA)	2	0.315	0.244
J. M. Torres-Moreno (LIA)	3	0.398	0.263
D. Buffoni (LIP6)	1	0.389	0.393
D. Buffoni (LIP6)	2	0.878	0.873
D. Buffoni (LIP6)	3	0.857	0.817
F. Rioult (GREYC)	1	0.725	0.717

FIG. 4 – F-scores pondérés ($\beta = 1$) pour toutes les soumissions ayant utilisé le score de confiance.

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

Les participants ont eu la possibilité d'associer un indice de confiance aux choix de la catégorie thématique ; l'identification du genre reposant sur deux choix (*Le Monde* ou Wikipédia), l'utilisation d'un indice de confiance pour cette sous-tâche n'était pas autorisé. Précisons que cet indice de confiance était proposé de manière optionnelle.

Sur les six participants du défi, trois ont utilisé l'indice de confiance pour pondérer leurs résultats. Sur ces trois participants, deux l'ont appliquée pour chaque soumission (J. M. Torres-Moreno au LIA et D. Buffoni au LIP6), l'autre équipe (F. Rioult au GREYC) ayant fait le choix de produire une soumission avec indices de confiance et deux soumissions sans indices de confiance.

L'utilisation de l'indice de confiance par les participants nous conduit a dresser trois constatations sur l'incidence de cet indice sur les résultats :

Incidence sur le classement des soumissions L'équipe ayant proposé des soumissions avec et sans utilisation de l'indice de confiance a obtenu de meilleurs résultats sur la soumission avec indice de confiance (en l'occurrence la première soumission).

Incidence sur le classement global des équipes Nous notons par ailleurs que deux des trois équipes arrivées dans les premières places de cette campagne (J. M. Torres-Moreno au LIA et D. Buffoni au LIP6) ont toutes deux utilisé les indices de confiance dans les résultats qui leur ont permis de se classer à ces niveaux.

Incidence sur le front de Pareto Enfin, nous remarquons que toutes les soumissions figurant sur le front de Pareto et la majorité des soumissions se trouvant à proximité immédiate de ce front (cf. graphiques n° 5, 6 et 7) sont également celles qui ont utilisé les indices de confiance dans les résultats.

Chaque participant ayant obtenu des résultats assez proches entre eux pour chaque sous-tâche considérée, il apparaît que le recours aux indices de confiance dans les résultats a permis de produire une différence sensible dans les classements finaux. La pondération des résultats document par document permet, *in fine*, de remonter l'ensemble des résultats d'une soumission donnée, par opposition aux soumissions qui n'utilisent pas ces indices de confiance et pour lesquelles les résultats présentés renvoient, pour chaque document, soit à une réussite (indice de confiance égal à 1), soit à un échec (indice de confiance égal à 0), sans que ces résultats ne soient pondérés par l'indice de confiance. Nous précisons que cette constatation ne porte que sur l'utilisation des indices de confiance dans les résultats soumis et que l'utilisation des outils et méthodes par chaque participant ne saurait être écartée dans la comparaison des résultats.

3 Front de Pareto

Définition Le front de Pareto est défini par l'ensemble des approches qui sont telles qu'aucune approche ne présente de meilleurs résultats pour tous les critères étudiés (rappel et précision dans le cas présent). Les approches qui ne sont pas sur le front de Pareto sont dites « dominées »[1].

Représentation graphique Le rappel est présenté sur l'axe des abscisses, la précision sur l'axe des ordonnées.

Le front de Pareto est symbolisé sur ces schémas par la ligne qui relie les meilleurs résultats entre eux. Les points sur le front de Pareto représentent les résultats qui, du point de vue du rappel, ou de la précision ou des deux à la fois, sont les meilleurs.

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
2	LIA : J. M. Torres-Moreno
3	LGI2P/LIRMM : M. Plantié
4	LIA : E. Charton, *équipe jeunes chercheurs*
6	LIP6 : D. Buffoni
8	LIFO/INaLCO : G. Cleuziou
10	GREYC : F. Rioult

L'ensemble des résultats étant présenté dans la tableau 1 d'une part, et les résultats étant assez groupés d'autre part, nous avons focalisé chaque graphique sur les nuages de points situés à proximité du front de Pareto. En conséquence, l'échelle utilisée diffère pour chaque graphique.

[1] `http://www.lri.fr/~aze/enseignements/bibs/2007-2008/docs/apprentissage-supervise.pdf`

3.1 Homogénéité des résultats

3.1.1 Identification du genre

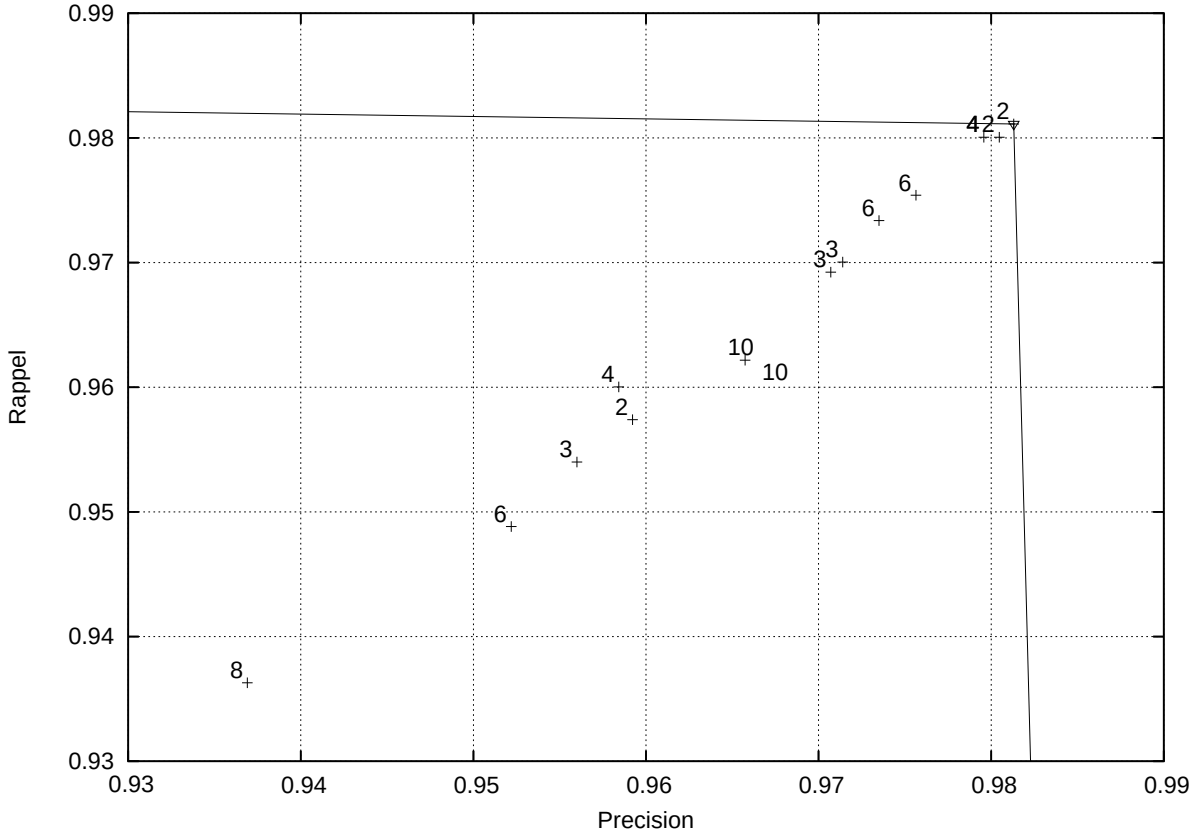

FIG. 5 – Front de Pareto pour la tâche 1, identification du genre.

La figure 5 montre un nuage de points homogène, les résultats des participants pour la tâche 1, identification du genre, sont très bons (pas un seul n'est inférieur à 0,856) au niveau du rappel et de la précision. En tête, les trois équipes ex-aequo, le LIA, le LIA jeunes chercheurs et le LIP6. Nous pouvons observer leur proximité avec le front de Pareto.

Nous relevons par ailleurs que l'ensemble des points sans exception aucune se situe sur une diagonale où rappel et précision sont en parfait équilibre, ce qui demeure assez exceptionnel et innatendu. Associant cette particularité aux très bons résultats obtenus par les participants, nous en déduisons que la tâche d'identification de genre avec seulement deux choix possibles s'est finalement révélée très facile. Les styles de l'encyclopédie Wikipédia et du journal *Le Monde* semblent bien se différencier.

3.1.2 Identification des catégories

Les points représentés sur la figure 6 sont plus disparates pour la tâche d'identification des catégories, mais avec également de bons résultats, la différence entre les deux figures pouvant également s'expliquer par le nombre de critères : 4 catégories à reconnaître contre deux genres distincts. Ici aussi, les équipes en tête sont les mêmes que pour l'identification du genre.

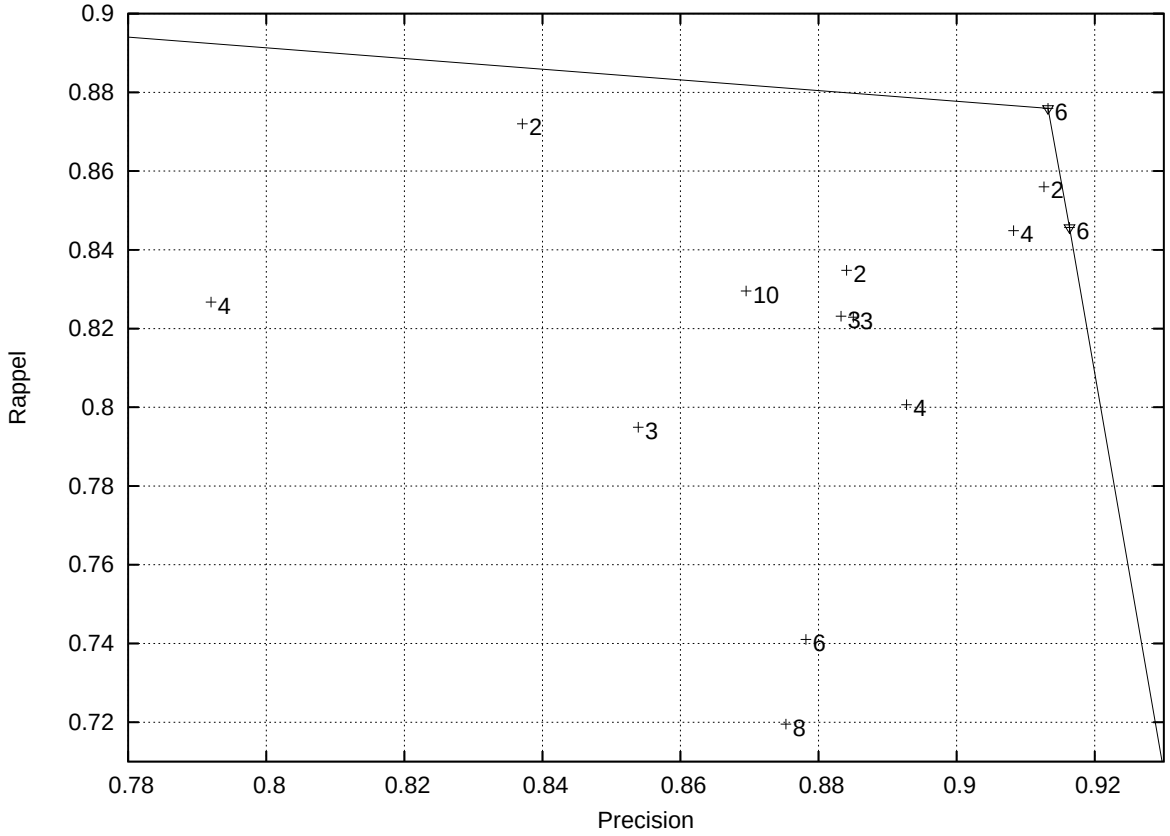

FIG. 6 – Front de Pareto pour la tâche 1, identification des catégories.

Au vu des résultats et contrairement à ce que nous avions imaginé lors de la préparation de ce défi, l'identification des catégories de la tâche 2 n'a pas posé plus de difficultés que celle de la tâche 1, alors que la tâche 2 comprend cinq catégories et que trois d'entre elles sont proches thématiquement (*France, International, Société*). Il apparaît même que la moitié des équipes a mieux réussi l'identification des catégories de la tâche n° 2 que celle de la tâche n° 1. Les résultats portant sur l'identification des catégories restent cependant très proches d'une tâche à l'autre comme le confirment les graphiques 2 et 3.

Il importe cependant de préciser que les soumissions les moins bonnes sont celles où l'identification des catégories de la tâche 2 a retourné de mauvais résultats (par exemple dans le cas des deuxièmes soumissions des équipes réunies autour de F. Rioult et d'E. Charton). Le plus grand nombre de catégories à identifier et la proximité sémantique de trois catégories sur cinq semble donc avoir occasionné quelques difficultés pour ces soumissions, la différence étant moins preignante pour l'identifcation des catégories de la tâche 1 sur ces mêmes soumissions.

4 Les méthodes utilisées par les participants

La confrontation de méthodes s'est avérée une fois de plus très productive, à la fois parce qu'elle montre des accords de performance sur des méthodes qui deviennent classiques, et qu'en même temps elle fait émerger des possibilités d'amélioration par des méthodes de conception plus originale.

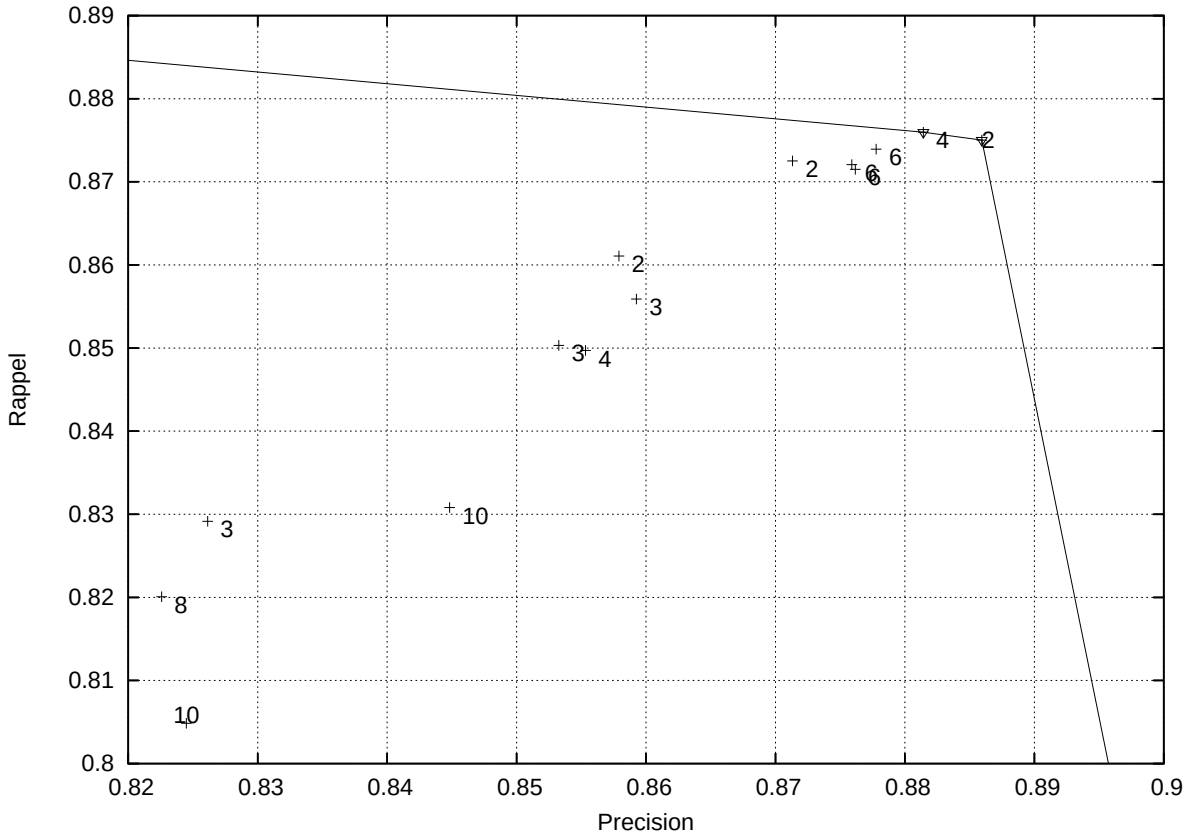

FIG. 7 – Front de Pareto pour la tâche 2, identification des catégories.

Le processus de classification comporte globalement deux étapes, en premier lieu une sélection des descripteurs du corpus, sur lesquels on applique ensuite un algorithme de classification. Plusieurs participants ont introduit une troisième étape de fusion des résultats de différents processus de classification ((Plantié *et al.*, 2008), (Béchet *et al.*, 2008), (Charton *et al.*, 2008)). Les méthodes de fusion améliorent généralement les résultats.

La sélection des descripteurs a donné lieu à une grande variété de méthodes. (Plantié *et al.*, 2008) et (Trinh *et al.*, 2008) ont choisi de ne pas lemmatiser les mots des documents. Chacun effectue ensuite une réduction de l'espace des mots, (Plantié *et al.*, 2008) par le calcul de l'information mutuelle sur chaque dimension, et (Trinh *et al.*, 2008) en calculant le score Okapi de chaque mot. (Béchet *et al.*, 2008) effectue des sélections par regroupement et normalisation de termes suivant des règles apprises automatiquement à partir de la maximisation du critère de discrimination. Des essais intéressants de classification du genre à partir du style et non du contenu thématique ont été proposés, soit à partir de la ponctuation des textes (Béchet *et al.*, 2008) ou des catégories morphosyntaxiques (Cleuziou & Poudat, 2008). Le critère d'impureté de Gini s'avère être un facteur discriminant efficace.

L'algorithme SVM a été plusieurs fois utilisé ((Cleuziou & Poudat, 2008), (Trinh *et al.*, 2008), (Plantié *et al.*, 2008), (Charton *et al.*, 2008)) avec de très bonnes performances. L'utilisation d'un noyau linéaire semble avoir été le meilleur choix. Cette méthode est très performante, mais sa limite est d'être peu explicative sur la discrimination linguistique qu'elle opère entre les classes. Par ailleurs, l'algorithme SVM séparant un corpus de documents en deux classes, des solutions doivent être trouvées pour résoudre le problème multi-classes. (Trinh *et al.*, 2008)

propose une méthode probabiliste efficace pour le choix de la meilleure classe, basée sur la maximisation de la log-vraisemblance conditionnelle. En revanche les classifieurs probabilistes à base de n-grammes de mots offrent l'avantage d'être explicatifs et permettent des observations intéressantes sur les discriminants linguistiques ((Charnois *et al.*, 2008), (Béchet *et al.*, 2008)). Les méthodes de boosting appliquées sur des classifieurs simples permettent d'en améliorer les performances.

Conclusion

Le défi cette année comportait plusieurs enjeux. Tout d'abord une possibilité de bilan sur les méthodes de classification en genre et de classification en thèmes, déjà largement explorées. Et ensuite une exploration de la classification thématique d'un mélange de genres, et de l'impact possible d'une classification en genre sur une classification en thème. Les systèmes rassemblant des méthodes innovantes de sélection des termes discriminants, des algorithmes éprouvés de classification, et des processus de fusion entre résultats, ont été les plus performants ((Béchet *et al.*, 2008), (Charton *et al.*, 2008)), en concurrence avec un système utilisant un SVM amélioré (Trinh *et al.*, 2008).

La classification en genre s'est révélée, pour les corpus choisis, plus facile. Les genres journalistiques et encyclopédiques se séparent bien, même dans des domaines semblables. Les catégories thématiques semblent plus difficiles à séparer, et la connaissance préalable du genre ne paraît produire qu'une amélioration à la marge. Il s'agissait d'un premier essai et d'autres possibilités de corpus sont à explorer, mettant en jeu par exemple des différences en genres spécialiste/néophyte dans des domaines de spécialité.

Références

BÉCHET F., EL-BÈZE M. & TORRES-MORENO J.-M. (2008). En finir avec la confusion des genres pour mieux séparer les thèmes. In *Actes TALN'08*, Avignon.

CHARNOIS T., DOUCET A., MATHET Y. & RIOULT F. (2008). Trois approches du GREYC pour la classification de textes. In *Actes TALN'08*, Avignon.

CHARTON E., CAMELIN N., ACUNA-AGOST R., GOTAB P., LAVALLEY R., KESSLER R. & FERNANDEZ S. (2008). Pré-traitements classiques ou par analyse distributionelle : application aux méthodes de classification automatique déployées pour DEFT08. In *Actes TALN'08*, Avignon.

CLEUZIOU G. & POUDAT C. (2008). Classification de textes en domaines et en genres en combinant morphosyntaxe et lexique. In *Actes TALN'08*, Avignon.

PAROUBEK P., BERTHELIN J.-B., EL AYARI S., GROUIN C., HEITZ T., HURAULT-PLANTET M., JARDINO M., KHALIS Z. & LASTES M. (2007). Résultats de l'édition 2007 du DÉfi Fouille de Textes. In *Actes de l'atelier de clôture du 3ème DÉfi Fouille de Textes*, p. 9–17, Grenoble : Association Française d'Intelligence Artificielle.

PLANTIÉ M., ROCHE M. & DRAY G. (2008). Défi DEFT08 : Classification de textes en genre et en thème : Votons utile ! In *Actes TALN'08*, Avignon.

TRINH A.-P., BUFFONI D. & GALLINARI P. (2008). Classifieur probabiliste avec Support Vector Machine (SVM) et Okapi. In *Actes TALN'08*, Avignon.