Préparation des données et analyse des résultats de DEFT'05

Érick Alphonse (1), Ahmed Amrani (2,3), Jérôme Azé (3),
Thomas Heitz (3), Amar-Djalil Mezaour (4), Mathieu Roche (3)

(1) MIG - INRA
Domaine de Vilvert, 78350 Jouy en Josas Cedex
Erick.Alphonse@jouy.inra.fr
(2) ESIEA Recherche
9 rue Vésale - 75005 Paris
amrani@esiea.fr
(3) Équipe IA, LRI - Université Paris-Sud
Bât. 490, 91405 Orsay Cedex
{amrani,aze,heitz,roche}@lri.fr
(4) Équipe IASI, LRI - Université Paris-Sud
Bât. 490, 91405 Orsay Cedex
mezaour@lri.fr

Résumé Le **DÉ**fi **F**ouille de **T**extes a consisté à supprimer les phrases non pertinentes dans un corpus de discours politiques en français. Il a eu lieu en 2005 et réuni onze équipes, totalisant une trentaine de participants. Cet article décrit les prétraitements effectués sur les corpus de F. Mitterrand et de J. Chirac dans le cadre de ce défi. Notamment, la conversion au format texte, le découpage en phrase, le classement des discours, l'introduction de phrases de F. Mitterrand dans les discours de J. Chirac et l'identification des dates et noms de personnes. Les résultats obtenus par les onze équipes participantes sont aussi présentés.

Abstract The text-mining challenge (DEFT) consisted of removing non relevant sentences from French corpora of political speeches. It took place in 2005 and brought together about thirty participants from eleven teams. This paper describes the preprocessings carried out on the corpora of F. Mitterrand and J. Chirac within the framework of this challenge. In particular, conversion to text format, sentence segmentation, classification of the speeches, introduction of F. Mitterrand's sentences into J. Chirac's speeches and identification of dates and people's names. The results obtained by the eleven participating teams are also presented.

1 Introduction

Le but du défi proposé consiste à supprimer les phrases non pertinentes dans un corpus de discours politiques en français. Ce défi porte le nom de DEFT (**DÉ**fi **F**ouille de **T**extes).

Ce défi, proche de la tâche *Novelty* du challenge TREC[1] (Soboroff, Harman, 2003; Amrani *et al.*, 2004), est motivé par le besoin de mettre en place des techniques de fouille de textes permettant soit d'identifier des phrases non pertinentes dans des textes, soit d'identifier des phrases particulièrement singulières dans des textes apparemment sans réel intérêt. Cette étape est préliminaire à tout processus d'extraction d'informations.

Par exemple, dans les corpus spécialisés (biologie, médecine, *etc.*) un travail conséquent est dédié à l'identification des phrases pertinentes pour ensuite y rechercher des informations spécifiques. Ce type de tâche consistant à effectuer un premier filtrage des textes est une étape préliminaire essentielle à effectuer pour la constitution de corpus pertinents et homogènes.

Ce type de prétraitements est aussi utilisé dans des tâches type *questions/réponses* (voir TREC).

Nous proposons ici une liste non exhaustive de tâches similaires à celle proposée dans DEFT'05 et pour lesquelles il devrait être possible de réutiliser avec peu de modifications les approches mises en œuvre pour répondre à DEFT'05.

- détection des passages les plus singuliers dans des textes quelconques (rupture de style, changement de contexte);
- détection de plagiats possibles dans des textes;
- détection des informations générales dans des corpus techniques.

Un corpus de textes, issus de la Présidence de Jacques Chirac (1995-2005), a été fourni aux participants de DEFT'05. Ce corpus est composé d'allocutions officielles du Président. Dans ce corpus, des passages issus d'un corpus d'allocutions du Président de la République François Mitterrand (1981-1995) sont insérés. Les passages d'allocutions de F. Mitterrand insérés sont composés d'au moins deux phrases successives. Chaque discours de J. Chirac contient zéro ou un passage extrait d'une allocution de F. Mitterrand.

Les passages de F. Mitterrand introduits traitent d'une thématique différente. Par exemple, dans les allocutions de J. Chirac évoquant la politique internationale, les phrases de F. Mitterrand introduites sont issues de discours traitant de politique nationale. Ainsi, la rupture thématique peut être une des manières de détecter les phrases issues du corpus de F. Mitterrand.

Cet article décrit plus spécifiquement les prétraitements effectués sur les corpus de F. Mitterrand et de J. Chirac.

La figure 1 illustre l'ensemble des traitements effectués pour DEFT'05.

[1] http://trec.nist.gov

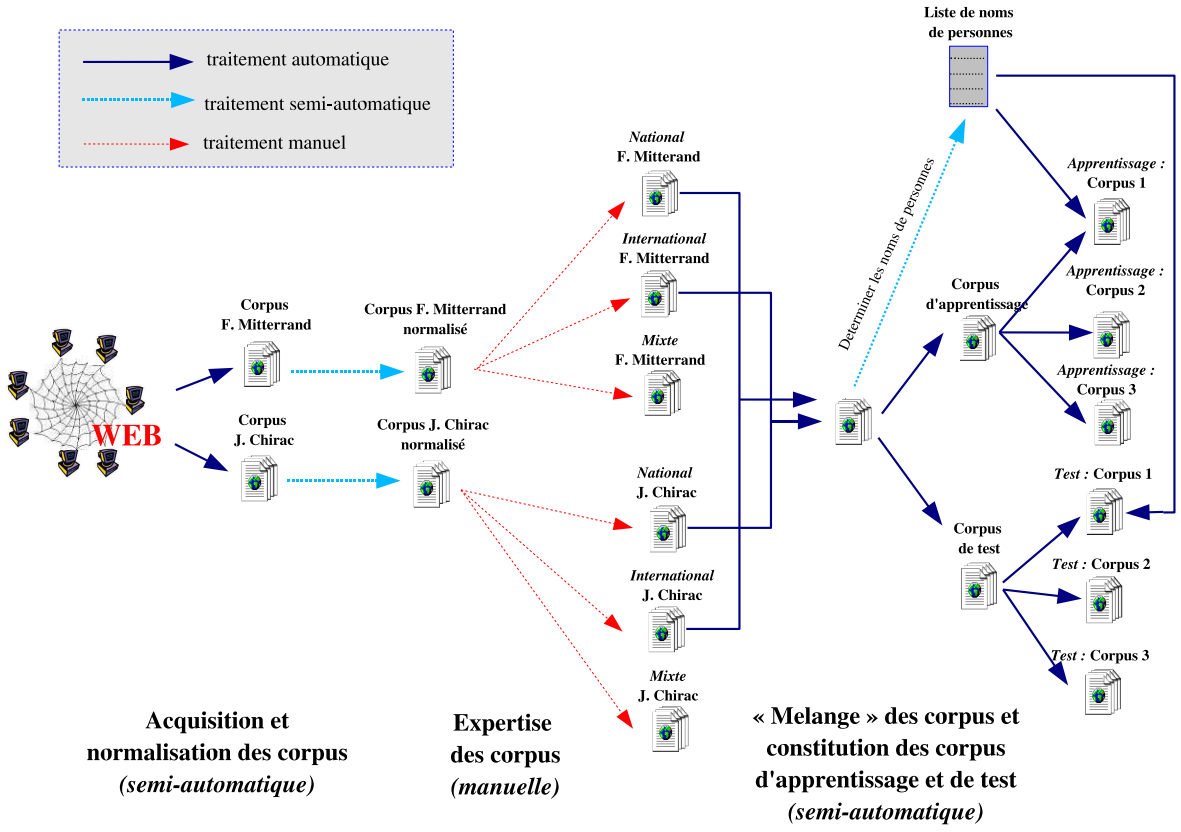

Figure 1: Chaîne globale de traitements de DEFT'05.

2 Acquisition des corpus

Les corpus composés des allocutions de J. Chirac et de F. Mitterrand ont été obtenus à partir des sites Web suivants :

- Corpus de J. Chirac (14 Mo sans considérer les balises HTML) : http://elysee.fr/
- Corpus de F. Mitterrand (17 Mo sans considérer les balises HTML) : http://discours-publics.ladocumentationfrancaise.fr/

3 Normalisation des corpus

Les corpus d'allocutions ont demandé un nombre de prétraitements important. Après avoir supprimé les commentaires et les balises HTML, les en-têtes des allocutions ont été enlevées (dates, lieux, *etc.*). Puis les entités au format SGML ont été transformées en caractères ISO8859-1. Par exemple, les entités « é » ont été remplacées par le caractère « é ».

Chacune des lignes des corpus fournis aux participants est composée d'une seule phrase. Pour identifier les phrases, il est nécessaire de repérer les ponctuations de fin de phrases (point final, point d'exclamation, point d'interrogation). Notons que comme dans les travaux de (Smadja, 1993), cette tâche nécessite le fait de ne pas considérer tous les points comme des ponctuations de fin de phrases (par exemple, les abréviations telles que R.M.I. pour Revenu Minimum d'Insertion ou M. pour Monsieur). De manière similaire aux travaux de (Rudolf, Świdziński, 2004), nous pouvons considérer que les points peuvent avoir des rôles spécifiques et sont utilisés dans différentes situations : abréviations, adresses internet, numéros de sections, *etc.*

Chaque locuteur peut utiliser régulièrement des phrases types du domaines qui pourraient permettre d'identifier les allocuteurs. À titre d'exemple, nous avons supprimé l'expression « Vive la République » qui est plus fréquente dans le corpus de F. Mitterrand (216 fois dans le corpus de F. Mitterrand contre 48 fois dans le corpus de J. Chirac).

Enfin, chaque phrase a été indexée rigoureusement grâce à une numérotation spécifique.

4 Expertise des corpus

Une étape d'expertise manuelle a alors été effectuée à partir des corpus normalisés. Le but de cette expertise a consisté à associer une catégorie à chacun des textes du corpus. Trois catégories ont été déterminées par le comité d'organisation (les auteurs de cet article) :

- Catégorie nationale
- Catégorie internationale
- Catégorie mixte ou ambiguë

Un discours traitant à 80% (estimation) d'une thématique déterminée sera associée à cette catégorie. Les discours contenant moins de 80% d'une thématique ont été associés à la catégorie mixte et ont été supprimés des données utilisées pour créer les corpus fournis aux participants.

Au total, 2523 textes ont été expertisés par les six organisateurs : 1200 allocutions de J. Chirac et 1323 allocutions de F. Mitterrand. Sur ces 2523 textes, 36.6% des textes ont été associés à la catégorie Nationale, 47.2% à la catégorie Internationale et 16,2% à la catégorie Mixte (voir tableau 1). Le détail complet des résultats donnés dans le tableau 1 montre notamment que les discours de F. Mitterrand ont davantage été associés à la catégorie Nationale que les allocutions de J. Chirac.

	F. Mitterrand	J. Chirac	**Global**
National	40.8%	31.9%	**36.6%**
International	45.0%	49.7%	**47.2%**
Mixte	14.1%	18.4%	**16.2%**

Table 1: Répartition des expertises par allocuteur.

Précisons que d'une période à l'autre, la répartition peut différer significativement. À titre d'exemples, les allocutions officielles de J. Chirac en 2002, année de l'élection présidentielle ont davantage été associées à la catégorie nationale (125 allocutions de J. Chirac en 2002 : 63 (50%) appartiennent à la catégorie Nationale, 46 (36.7%) appartiennent à la catégorie Internationale et 16 (12.8%) ont été associées à la catégorie Mixte).

5 Introduction des phrases de F. Mitterrand dans le corpus de J. Chirac

L'introduction des extraits de discours de F. Mitterrand dans les discours de J. Chirac a été réalisée en suivant les règles suivantes :

- Croisement des thématiques identifiées (politique nationale vs politique internationale)
- Sélection des extraits de discours de F. Mitterrand les plus "proches" des discours de J. Chirac pour l'introduction (voir paragraphe 5.2)
- Introduction d'au plus un passage de F. Mitterrand dans chaque discours de J. Chirac (voir paragraphe 5.3).

Le croisement des thématiques est lié à l'analyse présentée dans le tableau 1.

La distance entre un extrait de F. Mitterrand et un discours de J. Chirac est calculée en fonction des Ngrams de caractères et Ngrams mots. Nous avons calculé de manière systématique les Ngrams de caractères et de mots (pour n=1, 2 et 3) des discours de J. Chirac et des parties de discours de F. Mitterrand candidates à l'insertion (c-à-d. toutes les parties de discours sauf la première et la dernière).

Puis, nous avons comparé les discours de J. Chirac et parties de discours de F. Mitterrand (en tenant compte du croisement thématique) sur la base de ces Ngrams.

5.1 Comparaison des discours et des passages

Ces éléments sont comparés sur la base du score suivant :

$$score(d_C^{cat}, p_M^{\overline{cat}}) = score_{car}(d_C^{cat}, p_M^{\overline{cat}}) + score_{mot}(d_C^{cat}, p_M^{\overline{cat}})$$

avec

$$\begin{cases} cat & international\ ou\ national \\ \overline{cat} & catégorie\ opposée\ à\ cat \\ d_C^{cat} & discours\ de\ J.\ Chirac\ appartenant\ à\ cat \\ p_M^{cat} & partie\ de\ discours\ de\ F.\ Mitterrand\ appartenant\ à\ cat \end{cases}$$

$$score_{car}(d_C^{cat}, p_M^{\overline{cat}}) = \sum_{n=1}^{3} \left(\frac{1}{n}\right) \times 2 \times \frac{commun(d_C^{cat}, p_M^{\overline{cat}})}{|d_C^{cat}| + |p_M^{\overline{cat}}|}$$

où

$$\begin{cases} |x| & nombre\ de\ mots\ ou\ caractères\ de\ x \\ commun(d_C^{cat}, p_M^{\overline{cat}}) & nombre\ de\ mots\ ou\ caractères\ communs\ entre\ d_C^{cat}\ et\ p_M^{\overline{cat}} \end{cases}$$

$score_{mots}(d_C^{cat}, p_M^{\overline{cat}})$ est calculé selon la même formule mais sur la base des Ngrams[2] entre mots et non pas entre caractères.

5.2 Sélection des passages à insérer

Ayant calculé ce score pour tous les couples possibles $(d_C^{cat}, p_M^{\overline{cat}})$, nous retenons pour chaque d_C^{cat} les vingt "meilleurs" $p_M^{\overline{cat}}$ (c-à-d. tels que $score(d_C^{cat}, p_M^{\overline{cat}})$ soient les plus élevés). Ces vingt candidats à l'insertion sont triés par valeurs décroissantes du score.

Puis, les discours de J. Chirac sont parcourus aléatoirement et les insertions de passages de discours de F. Mitterrand sont réalisées de la manière suivantes :

Soient d_C^{cat} le discours de J. Chirac étudié et $\mathcal{L}_{p_M^{\overline{cat}}}^{d_C^{cat}}$ la liste des passages candidats à l'insertion. Soit $\mathcal{E}_{p_M^{\overline{cat}}}$ l'ensemble des passages de discours de F. Mitterrand déjà introduits dans des discours de J. Chirac.

La liste ordonnée $\mathcal{L}_{p_M^{\overline{cat}}}^{d_C^{cat}}$ est parcourue depuis le premier passage vers le dernier jusqu'à trouver un passage qui soit absent de $\mathcal{E}_{p_M^{\overline{cat}}}$. Si un tel passage existe, il est introduit dans d_C^{cat}, puis dans $\mathcal{E}_{p_M^{\overline{cat}}}$. Par contre, si aucun passage n'est trouvé alors le discours de J. Chirac étudié n'est pas modifié (c-à-d. le discours est donc "non bruité").

5.3 Insertion d'un passage de F. Mitterrand dans un discours de J. Chirac

La position d'un passage à insérer est déterminée en respectant les contraintes suivantes :

- ni avant le premier, ni après le dernier paragraphe[3] du discours de J. Chirac.

[2] L'outil **nsp-v0.71** a été utilisé pour calculer les Ngrams (`http://www.d.umn.edu/~tpederse/nsp.html`).

[3] Un paragraphe correspond à un bloc de texte entre balises HTML <p> ou séparé par deux balises
.

- aléatoirement dans le reste du discours et entre deux paragraphes

Le corpus ainsi constitué a été divisé en deux sous-ensembles : le corpus d'apprentissage et le corpus de test. Nous avons utilisé 70% des discours pour constituer le corpus d'apprentissage et les 30% restant pour le test. Les discours ont été choisis de manière aléatoire et stratifiée. En effet, nous avons garanti par construction que les proportions de discours "bruités" et "non bruités", dans les corpus de test et d'entraînement, sont identiques à celles observées dans le corpus initial.

5.4 Remarque

Il peut arriver que deux thématiques identiques (Nationale ou Internationale) soient insérées dans un même texte. Ceci peut s'expliquer par le fait qu'un texte de F. Mitterrand associé à une catégorie Nationale (resp. Internationale) peut comporter des passages d'une catégorie Internationale (resp. Nationale). Ces passages de F. Mitterrand de la catégorie Internationale (resp. Nationale) bien que minoritaires dans l'allocution associée à la catégorie Nationale (resp. Internationale) pourraient alors être introduits dans une allocution de J. Chirac de la catégorie Internationale (resp. Nationale).

6 Constitution des trois corpus avec et sans informations relatives aux noms de personnes et aux années

Nous rappelons que le défi DEFT'05 comporte trois tâches distinctes :

- **Tâche 1 :** Identifier les phrases de F. Mitterrand dans le corpus de test numéro 1 (corpus ne comportant ni années, ni noms de personnes).
- **Tâche 2 :** Identifier les phrases de F. Mitterrand dans le corpus de test numéro 2 (corpus ne comportant pas d'années).
- **Tâche 3 :** Identifier les phrases de F. Mitterrand dans le corpus de test numéro 3 (corpus avec la présence des années et des noms de personnes).

Pour constituer les corpus 1 et 2, nous avons du identifier les dates (années) ainsi que les noms de personnes. Ces identifications sont détaillées ci-dessous.

6.1 Identification des dates

Seules les années situées dans l'intervalle $[1900 : 2099]$ ont été identifiées. Ces années pourraient en effet faciliter l'identification des phrases issues du corpus de F. Mitterrand.

Ainsi les années de la forme 19xx et 20xx où « x » est un chiffre quelconque ont été identifiées et remplacées par une balise <date>. De même, les intervalles entre années ont été reconnus : 19xx-19xx, 19xx-20xx, 20xx-20xx et xx-xx.

Chacune des dates de ces intervalles ont également été remplacées par une balise <date>.

Les dates au format "1er février 2004" n'ont pas été identifiée et peuvent donc figurer dans les corpus, sous la forme "1er février <date>"

Ce traitement a permis de constituer les corpus utiles pour les tâches 1 et 2.

6.2 Identification des noms de personnes

Une liste de noms de personnes a dû être établie manuellement. Les membres du Comité d'Organisation de DEFT'05 ont ainsi analysés les suites de mots suivants afin d'identifier les noms de personnes :

- couples de mots commençant par une majuscule.
- couples de mots commençant par une majuscule avec une particule intercalée entre les deux mots.
- particule suivi d'un mot en majuscules.

Les particules utilisées (avec et sans majuscules) sont les suivantes : Abd, Al, Ap, Ben, Bin, D', Da, Dalle, Dall', Dell', De, De La, De Los, Del, Dela, Della, Delle, Den, Der, Di, Du, El, Ibn, La, Le, Li, Lo, Mac, Mc, O', Of, Saint, San, Van, Van Den, Van Der, Von, Von Der, y.

De plus, un dictionnaire de noms de personnes composés d'un seul mot a été constitué (par exemple, Picasso, Dali, *etc.*).

Les noms de personnes étant identifiés, nous les avons remplacés par une balise <nom>.

Ce traitement a permis de constituer le corpus utile pour la tâche 1.

7 Traitement final des corpus

Le dernier traitement a consisté à maintenir en majuscule seulement la première lettre des noms de personnes. En effet, dans le corpus de J. Chirac la plupart des noms de personnes sont écrits en majuscules (Jacques CHIRAC, François MITTERRAND, *etc.*). Ainsi, l'identification des noms en majuscules aurait pu être une règle simple mais efficace pour reconnaître les phrases issues du corpus de F. Mitterrand et de J. Chirac des tâches 1 et 2. Pour corriger cette situation, nous avons uniformisé l'écriture des noms de personnes en écrivant seulement en majuscule la première lettre du nom de personne : MITTERRAND → Mitterrand. Bien entendu les acronymes (PS, RPR, EDF, *etc.*) sont maintenus en majuscules. Certains noms de personnes écrits en majuscules contiennent moins d'informations qu'un même nom écrit en majuscules. En effet, dans des cas relativement nombreux, les noms en majuscules ne comportent pas d'accents (par exemple « JUPPE » qui correspond en lettres minuscules à « Juppé »). Les accents doivent donc être restitués lors du passage majuscules/minuscules. Une manière semi-automatique de procéder consiste à relever la présence des noms de personnes (nom commençant par une majuscule) que l'on trouve dans le texte avec des accents. Dans ce cas, nous pouvons décider d'apposer par défaut l'accent omis. Si aucun mot similaire (avec accents) n'est repéré dans le corpus, et sans utiliser de ressources extérieures, il est nécessaire d'expertiser ces

	Corpus d'apprentissage	Corpus de test
Taille moyenne des phrases successives de F. Mitterrand	18.8	19.1
Pourcentage d'allocutions sans phrases de F. Mitterrand insérées	31.9% (187/587)	32.3% (95/294)
Nombre d'étiquettes <nom> par rapport au nombre de mots du corpus	0.18% (2511/1420833)	0.21% (1331/616584)
Nombre d'étiquettes <date> par rapport au nombre de mots du corpus	0.13% (1846/1420833)	0.12% (774/616584)

Table 2: Comparaison des corpus d'apprentissage et de test.

noms et d'y apposer manuellement les accents manquants.

8 Similarités entre les corpus d'apprentissage et de test

Les corpus d'apprentissage et de test ont été constitués simultanément, c'est la raison pour laquelle ils ont des caractéristiques similaires. Ainsi, les méthodes mises en œuvre sur les corpus d'apprentissage peuvent être appliquées sur les corpus de test sans nécessiter d'adaptations spécifiques. Nous donnons dans le tableau 2 les caractéristiques essentielles des corpus d'apprentissage et de test.

Remarquons que le pourcentage d'étiquettes <nom> dans les corpus de test est plus élevé que dans les corpus d'apprentissage (voir tableau 2). Cela peut s'expliquer par le fait que nous avons apporté une attention toute particulière à la préparation des corpus de test pour lesquels les participants avaient seulement deux à quatre jours de traitements possibles.

9 Résultats obtenus par les équipes participantes

Onze équipes ont participé à DEFT'05. Ces onze équipes sont issues de neuf laboratoires différents et totalisent une trentaine de participants.

Les résultats obtenus sont assez variés (en termes de précision, rappel et Fscore) et tendent donc à montrer que les tâches à réaliser étaient non triviales, tout en restant faisables. Nous rappelons que toutes les exécutions ont été évaluées en calculant le F_{score} (avec $\beta = 1$, voir formule (1)).

$$F_{score}(\beta) = \frac{(\beta^2 + 1) \times Précision \times Rappel}{\beta^2 \times Précision + Rappel} \quad (1)$$

Le tableau 3 présente les Fscores obtenus par les différentes équipes pour chaque tâche (nous avons présenté les Fscores moyens calculés à partir des différentes exécutions soumises). Ces résultats sont triés par Fscore décroissant sur la base de la première tâche. Les résultats indiqués en italiques correspondent à des équipes n'ayant soumis qu'une seule exécution pour la tâche concernée.

	tâche 1	tâche 2	tâche 3
équipe 1	0.870239	0.884376	0.880458
équipe 2	0.860471	0.851721	0.866254
équipe 3	0.819636	0.821018	0.819075
équipe 4	0.759816	0.741895	0.745863
équipe 5	0.751278	0.754767	0.755094
équipe 6	0.731591	0.793889	0.788093
équipe 7	0.561811	0.559366	0.573122
équipe 8	0.493809	0.521142	0.507484
équipe 9	0.49241	0.560066	0.563089
équipe 10	0.325341	0.306647	0.305337
équipe 11	0.176951	0.176951	0.41729

Table 3: Meilleurs Fscores des différents équipes pour chaque tâche.

L'analyse des résultats détaillés (par exécution) sur la base du Fscore avec $\beta = 1$ permet de voir que la plupart des équipes ont amélioré leurs résultats au fur et à mesure des tâches (voir les courbes 2). Ainsi, l'ajout d'informations (noms de personnes, puis années) représente une aide réelle pour les différents systèmes représentés dans ce défi.

De plus, l'analyse du front de Pareto associé à la tâche 1 (voir Figure 3) montre que plusieurs équipes se trouvent sur le front de Pareto et donc qu'en fonction de la valeur de β choisie, l'ordre des approches en fonction du Fscore peut être modifié. Nous obtenons les mêmes résultats pour les tâches 2 et 3.

10 Conclusion

La problématique abordée dans DEFT'05 est relative à une tâche importante dans tout processus de fouille de données et constitue une étape préliminaire aux phases d'extraction d'informations.

L'implication de nombreuses équipes de recherche dans ce défi montre l'intérêt réel de la communauté pour ce problème et notamment pour la comparaison et l'évaluation de différentes méthodes de prétraitement des données et d'extraction d'informations.

La diversité des résultats obtenus par les équipes ayant participé montre que cette tâche représente une réelle difficulté pour la communauté et l'un des avantages de DEFT'05 est lié à la nature artificielle du corpus qui permet ainsi une évaluation plus objective des résultats obtenus par les différentes équipes.

L'engouement de la communauté pour ce défi et les différentes propositions d'extensions de DEFT vont permettre la poursuite de ce défi l'année prochaine. Les tâches précises restent à déterminer mais DEFT'05 a réussi à fédérer la communauté francophone de fouille de textes. Une extension envisageable et intéressante serait liée à l'étude de données réelles et à la définition d'un nouveau problème pour DEFT'06.

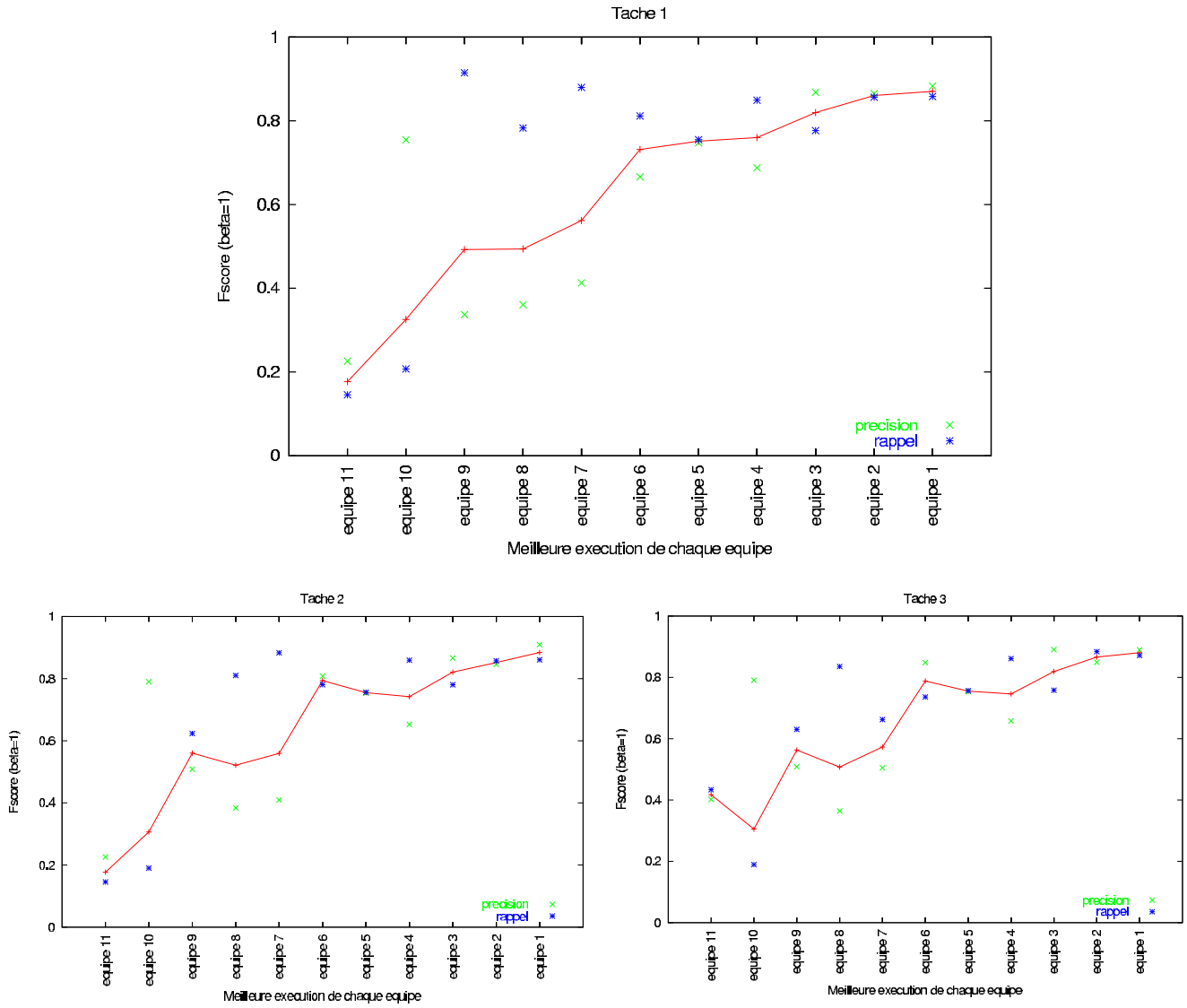

Figure 2: Fscore ($\beta = 1$) pour les meilleurs exécutions.

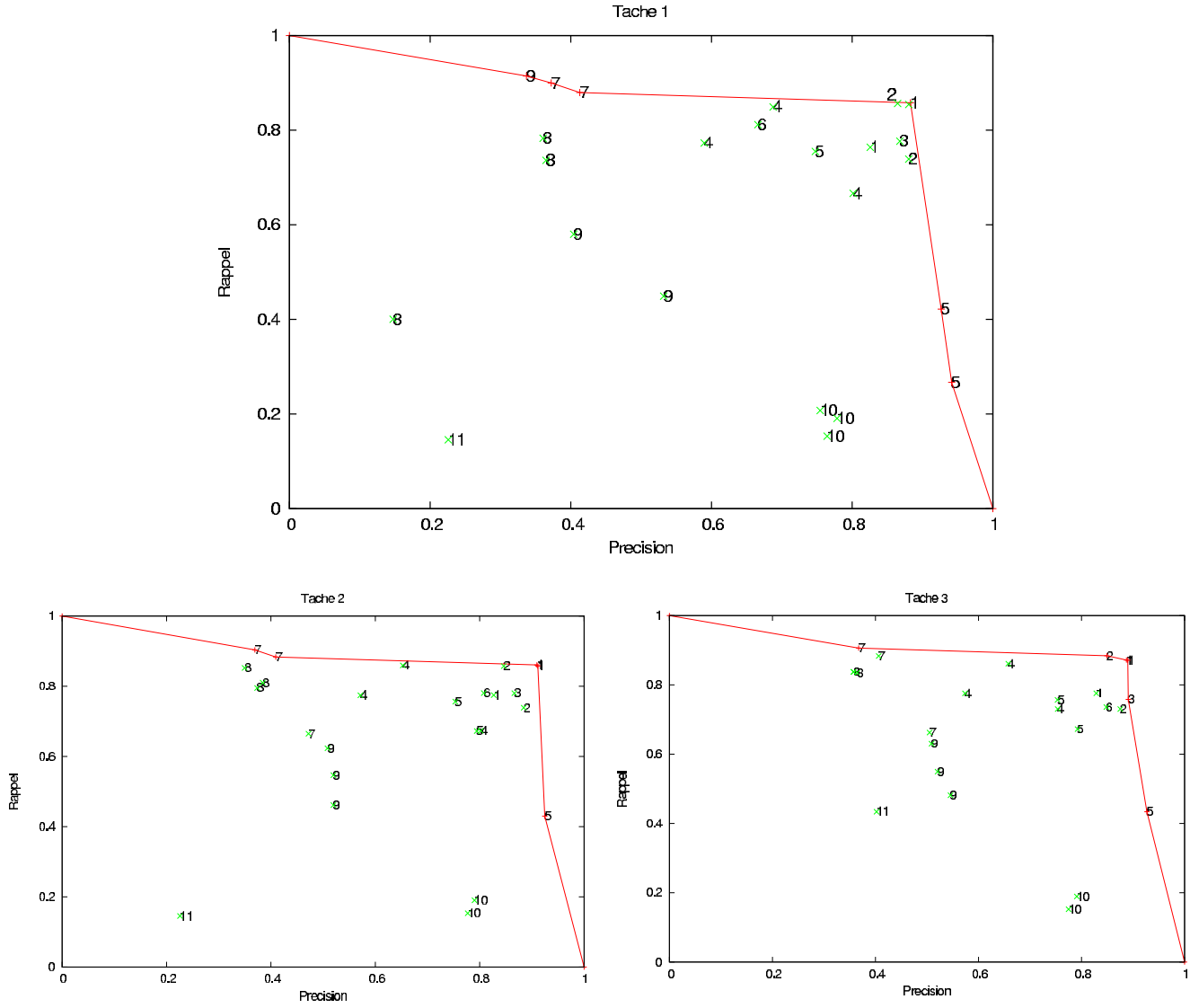

Figure 3: Front de Pareto pour les tâches 1, 2 et 3.

Références

Rudolf M., M. Świdziński (2004), Automatic utterance boundaries recognition in large Polish text corpora, *Proceedings of IIPWM'04 (Intelligent Information Processing and Web Mining), Springer Verlag series "Advances in Soft Computing"*, 247-256.

Smadja F. (1993), Retrieving collocations from text: Xtract, *Computational Linguistics*, Vol. 191, p143-177.

Soboroff I., Harman D. (2003), Overview of the TREC 2003 Novelty Track, *NIST Special Publication: SP 500-255 The Twelfth Text Retrieval Conference (TREC 2003).*

Amrani. A, Azé. J, Heitz. T, Kodratoff. Y and Roche. M (2004), From the texts to the concepts they contain: a chain of linguistic treatments, *Proceedings of TREC'04 (Text REtrieval Conference), National Institute of Standards and Technology, Gaithersburg Maryland USA*, pages 712-722.