Bilan du Premier Défi Francophone de Fouille de Textes

Jérôme Azé*, Mathieu Roche**,
Érick Alphonse***, Ahmed Amrani*,****
Thomas Heitz*, Amar-Djalil Mezaour*****

* LRI – Université Paris-Sud
Bât. 490, 91405 Orsay Cedex
{aze,amrani,heitz}@lri.fr
**LIRMM – Université de Montpellier 2
161 rue Ada, 34392 Montpellier Cedex 5
mroche@lirmm.fr
***LIPN – Université de Paris Nord
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
Erick.Alphonse@lipn.univ-paris13.fr
****ESIEA Recherche
9 rue Vésale, 75005 Paris
amrani@esiea.fr
*****Exalead
10 place de la Madeleine, 75019 Paris
Amar-Djalil.Mezaour@exalead.com

Résumé. Le **DÉ**fi **F**ouille de **T**extes (DEFT) 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 phrases, 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.

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 pour **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.

[1] Text REtrieval Conferences : http ://trec.nist.gov

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.

Le cadre d'étude de l'attribution d'auteur pour un segment de texte ne semble pas très normalisé malgré les nombreuses études qui existent [Rudman, 1997]. C'est en partie pour donner un cadre d'étude standardisé que ce défi a été mis en place. Nous partons de l'hypothèse que les auteurs laissent une empreinte dans les textes décelable à l'aide de méthodes d'analyse statistique [Baayen *et al.*, 2002].

Un corpus de textes, issus de la Présidence de J. Chirac (1995-2005), a été fourni aux participants de DEFT'05. Ce corpus obtenu à partir du site http ://elysee.fr/, d'une taille de 14 Mo de texte brut, est composé d'allocutions officielles du Président. Dans ce corpus, des passages issus d'un corpus d'allocutions du Président de la République F. 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. Le corpus de discours de F. Mitterrand représentant 17 Mo de texte brut a été obtenu à partir du site http ://discours-publics.ladocumentationfrancaise.fr/.

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.

Le défi DEFT'05 se présente sous la forme de trois tâches distinctes à résoudre :

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

Ces trois tâches sont données dans un ordre décroissant de difficulté car l'année et/ou les noms de personnes peuvent être de bons indicateurs des périodes de présidence.

L'article présente toute la chaîne de traitements du défi (de la préparation des données jusqu'à l'analyse des résultats de DEFT'05).

Une large part de cet article traite de la préparation des données de DEFT (section 2) effectuée par les auteurs. Le but était non seulement de normaliser les corpus pour les trois tâches mais également d'introduire les phrases de F. Mitterrand dans les discours de J. Chirac de manière pertinente afin de rendre le défi abordable sans être trivial.

La section 3 présente une comparaison succincte des approches utilisées par les participants de DEFT'05 (méthodes de classification, apports linguistiques, *etc.*). L'objet de cet article n'est pas de décrire de manière détaillée les méthodologies choisies par les participants. Les approches des participants sont présentées dans les autres articles de cette revue.

La section 4 analyse les résultats obtenus par les participants. Cette analyse s'appuie sur le calcul du F_{score}, critère d'évaluation fixé pour le défi et pour la plupart des compétitions internationales de fouille de textes. De plus, une discussion fondée sur l'utilisation du front de Pareto complète l'analyse des résultats.

Enfin, pour conclure, nous présentons une discussion qui nous a permis de motiver la mise en œuvre du défi DEFT'06 en septembre 2006.

2 Préparation des données

La figure 1 illustre l'ensemble de la chaîne de préparation des données du défi. Toutes les étapes de cette chaîne sont décrites dans les sections suivantes. Ainsi, après l'acquisition des corpus utilisés pour de défi, les différents traitements présentés dans cette section sont relatifs à la normalisation des corpus, leur expertise, la méthode utilisée pour introduire les phrases de F. Mitterrand dans le corpus de J. Chirac, la préparation des corpus pour les trois tâches du défi, les derniers traitements nécessaires en fin de chaîne. Une dernière section (section 2.6) présente une comparaison entre les corpus de test et d'apprentissage constitués.

2.1 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

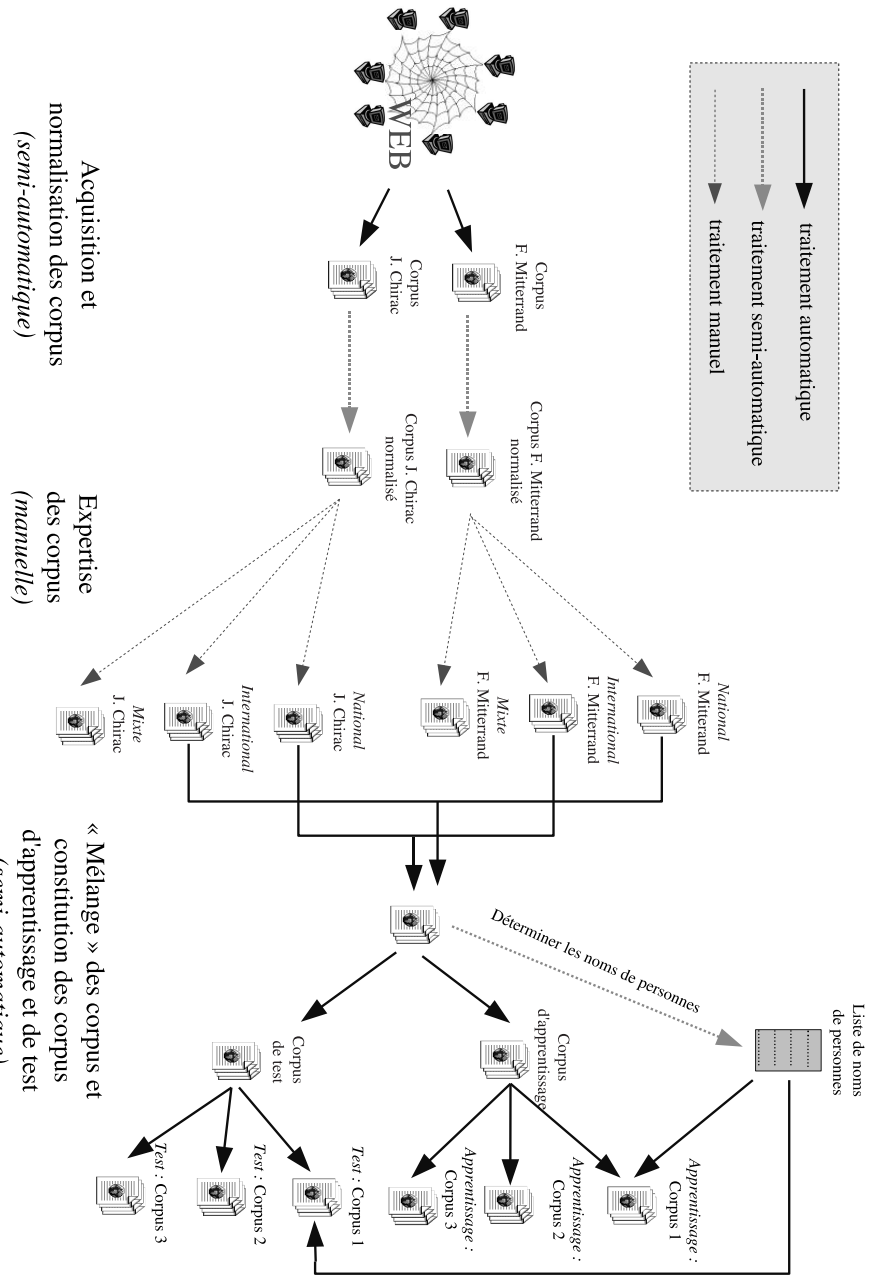

FIG. 1 – Chaîne globale de traitements de DEFT'05.

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 domaine 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 grâce à une numérotation spécifique.

2.2 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 et auteurs de cet article : 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**
Nationale	40.8%	31.9%	**36.6%**
Internationale	45.0%	49.7%	**47.2%**
Mixte	14.1%	18.4%	**16.2%**

TAB. 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. Sur les 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.

L'expertise présentée dans cette section sera utilisée pour introduire les phrases de F. Mitterrand dans le corpus de J. Chirac (voir section suivante).

2.3 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 respectant les points suivants :

- 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 ;
- introduction d'au plus un passage de F. Mitterrand dans chaque discours de J. Chirac.

Le croisement des thématiques est lié à l'analyse présentée dans le tableau 1. Nous présentons le point 2 et le point 3 respectivement dans les deux paragraphes suivants.

2.3.1 Sélection des passages à insérer

La sélection des passages se fait sur la base d'un score de similarité entre un extrait de F. Mitterrand et un discours de J. Chirac. Ce calcul est fonction des n-grammes[2] de caractères et n-grammes de mots (n caractères ou mots consécutifs présents dans les textes permettant ainsi de les caractériser).

Calcul du score.

Nous avons calculé de manière systématique les n-grammes 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'est-à-dire 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 n-grammes.

Ces éléments sont comparés sur la base du score suivant :

$$score(d_C^{cat}, p_M^{\overline{cat}}) = score_{car}(d_C^{cat}, p_M^{\overline{cat}}) + score_{mot}(d_C^{cat}, p_M^{\overline{cat}}) \quad (1)$$

avec

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

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

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

où

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

$score_{mot}(d_C^{cat}, p_M^{\overline{cat}})$ est calculé selon la même formule mais sur la base des n-grammes entre mots et non pas entre caractères.

Prenons un exemple de trois fragments de phrases d_1, d_2 et d_3 et calculons les scores obtenus. Dans cet exemple, nous considérons ici que la phrase d_1 appartient à un discours de François Mitterrand, catégorie Nationale et que les phrases d_2 et d_3 appartiennent à un discours de Jacques Chirac, catégorie Internationale.

Exemples	
d_1 :	`Le ministre Jospin`[a]
d_2 :	`Le premier ministre Juppé`[b]
d_3 :	`Le premier ministre Jospin`[c]

[a] Ministre de l'Éducation Nationale (1988-1992) sous la Présidence de F. Mitterrand
[b] Premier Ministre (1995-1997) sous la Présidence de J. Chirac
[c] Premier Ministre (1997-2002) sous la Présidence de J. Chirac

Nous calculons ci-dessous les scores entre les phrases d_2 et d_1 puis entre d_3 et d_1 qui respectent le croisement thématique et des locuteurs. Le détail des n-grammes est donné en annexe dans le tableau 10.

Calcul du score fondé sur les n-grammes de caractères – formule (2)
$commun_{n=1}(d_2, d_1) = 11$, $commun_{n=1}(d_3, d_1) = 12$
$commun_{n=2}(d_2, d_1) = 11$, $commun_{n=2}(d_3, d_1) = 15$
$commun_{n=3}(d_2, d_1) = 10$, $commun_{n=3}(d_3, d_1) = 15$
$\lvert d_1\rvert_{n=1} = 12$, $\lvert d_2\rvert_{n=1} = 13$, $\lvert d_3\rvert_{n=1} = 12$
$\lvert d_1\rvert_{n=2} = 15$, $\lvert d_2\rvert_{n=2} = 21$, $\lvert d_3\rvert_{n=2} = 21$
$\lvert d_1\rvert_{n=3} = 16$, $\lvert d_2\rvert_{n=3} = 23$, $\lvert d_3\rvert_{n=3} = 24$
$score_{car}(d_2, d_1) = \binom{2}{1} \times \frac{11}{13+12} + \binom{2}{2} \times \frac{11}{21+15} + \binom{2}{3} \times \frac{10}{23+16} = 1.36$
$score_{car}(d_3, d_1) = \binom{2}{1} \times \frac{12}{12+12} + \binom{2}{2} \times \frac{15}{21+15} + \binom{2}{3} \times \frac{15}{24+16} = 1.67$

Calcul du score fondé sur les n-grammes de mots – formule (2)
$commun_{n=1}(d_2, d_1) = 1$, $commun_{n=1}(d_3, d_1) = 2$
$commun_{n=2}(d_2, d_1) = 0$, $commun_{n=2}(d_3, d_1) = 1$
$commun_{n=3}(d_2, d_1) = 0$, $commun_{n=3}(d_3, d_1) = 0$
$\lvert d_1\rvert_{n=1} = 2$, $\lvert d_2\rvert_{n=1} = 3$, $\lvert d_3\rvert_{n=1} = 3$
$\lvert d_1\rvert_{n=2} = 2$, $\lvert d_2\rvert_{n=2} = 3$, $\lvert d_3\rvert_{n=2} = 3$
$\lvert d_1\rvert_{n=3} = 1$, $\lvert d_2\rvert_{n=3} = 2$, $\lvert d_3\rvert_{n=3} = 2$
$score_{mot}(d_2, d_1) = \binom{2}{1} \times \frac{1}{3+2} + \binom{2}{2} \times \frac{0}{3+2} + \binom{2}{3} \times \frac{0}{2+1} = 0.4$
$score_{mot}(d_3, d_1) = \binom{2}{1} \times \frac{2}{3+2} + \binom{2}{2} \times \frac{1}{3+2} + \binom{2}{3} \times \frac{0}{2+1} = 1$

Calcul du score fondé sur les n-grammes de caractères et de mots – formule (1)
$score(d_2, d_1) = 1.36 + 0.4 = 1.76$
$score(d_3, d_1) = 1.67 + 1 = 2.67$

Le résultat précédent montre que les phrases d_1 et d_3 respectant le croisement thématique et de locuteurs différents sont, d'après le score établi, plus proches que les phrases d_1 et d_2.

Choix des passages à insérer fondé sur les scores calculés.

Afin que les parties des discours de F. Mitterrand soient intégrées dans le corpus de J. Chirac sans qu'il y ait de ruptures trop évidentes, nous avons utilisé les scores calculés avec la formule 1 pour choisir les passages à insérer. Le but était alors d'introduire les passages des discours de F. Mitterrand dans les discours de J. Chirac ayant des scores élevés. Avec l'exemple précédent, les discours d_1 et d_3 sont alors sélectionnés.

Plus formellement et plus généralement la méthode de sélection des passages à insérer respecte le principe suivant. Ayant calculé le 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'est-à-dire 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 suivante :

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'est-à-dire que le discours est donc "non bruité").

2.3.2 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 ;
- 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 : nous avons garanti par construction que les proportions

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

de discours "bruités" et "non bruités", dans les corpus de test et d'apprentissage, sont identiques à celles observées dans le corpus initial.

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).

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

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

2.4.1 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 identifiées 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.

2.4.2 Identification des noms de personnes

Pour constituer le corpus de la tâche 1, chaque nom de personne repéré dans une liste a été remplacé dans le document par une balise <nom>. Cette liste a été établie semi-automatiquement.

Tout d'abord, un dictionnaire de noms de personnes composés d'un seul mot a été constitué (par exemple, Picasso, Dali, *etc.*). Ensuite, les membres du Comité d'Organisation de DEFT'05 ont analysé les suites de mots suivants dans l'ensemble des corpus 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.

2.5 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.
Lors du passage majuscules/minuscules, il est très fréquent que les accents soient à restituer (par exemple « JUPPE » correspond en lettres minuscules à « Juppé »). 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 noms et d'y apposer manuellement les accents manquants.

2.6 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

	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)

TAB. 2 – Comparaison des corpus d'apprentissage et de test.

traitements possibles.

Après avoir décrit les traitements effectués pour la préparation du corpus de DEFT-'05, la section suivante résume les différentes méthodes utilisées par les participants. Des descriptions plus précises des approches utilisées sont développées dans les autres articles de cette revue.

3 Approches utilisées pour la résolution des tâches de DEFT'05

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

Le tableau 3 présente les caractéristiques des traitements effectués par les différentes équipes. Le prétraitement le plus efficace semble être les n-grammes de mots compris entre 1 et 4 (équipes 1, 3, 4 et 7-9), tandis que la suppression des mots (2, 4, 8 et 10), tels les mots vides, ne semble pas avoir une influence déterminante sur le F_{score}.

Les méthodes de classification les plus employées ont été les chaînes de Markov (équipes 1, 2, 4, 6 et 7-9) et l'apprentissage bayésien (1, 3, 4 et 8). Les meilleurs F_{scores} ont été obtenus grâce à la combinaison de ces deux méthodes (1 et 4) bien que les chaînes de Markov seules semblent déjà très efficaces (2).

L'apport linguistique le plus utilisé est l'étiquetage grammatical (équipes 1, 6, 7-9, 8 et 10) même si les résultats sont très différents d'une équipe à l'autre. L'utilisation d'entités nommées (1 et 6) semble avoir été efficace pour l'équipe 1 grâce à l'utilisation d'un dictionnaire conséquent. Les approches à base de vecteurs de termes ne semblent pas les plus adaptées (8 et 10).

Les participants à DEFT'05 ont travaillé durant près de deux mois sur les corpus

Numéro d'équipe	**1**	**2**	3	**4**	**5**	6	**7-9**	8	**10**
Types de prétraitements									
Suppression de mots		✓		✓				✓	✓
N-lettres			7						
N-mots	1-2		4	1-3			4		
Méthodes de classification									
Chaînes de Markov	✓	✓		✓		✓	✓		
Viterbi	✓	✓		✓			✓		
Bayes	✓		✓	✓				✓	
SVM				✓		✓			
Apports linguistiques									
Vecteurs de termes								✓	✓
Étiquetage grammatical	✓					✓	✓	✓	✓
Relations syntaxiques					✓				✓
Entités nommées	✓					✓			

TAB. 3 – Comparaison des traitements effectués par les différentes équipes. Les **équipes qui ont un article dans cette revue** sont les équipes numérotées 1, 2, 4, 5, 7-9 et 10.

d'apprentissage (étiquetés selon les locuteurs). Les corpus de test (sans l'information sur les locuteurs) ont ensuite été mis à disposition. Les résultats obtenus avec les approches résumées dans cette section ont alors été envoyés au comité d'organisation. L'analyse de ces résultats à partir des corpus de test est présentée dans la section suivante.

4 Analyse des résultats de DEFT'05

Les résultats obtenus sont assez variés (en termes de précision, rappel et F_{score}) et tendent donc à montrer que les tâches à réaliser étaient non triviales, tout en restant parfaitement abordables.

4.1 Analyse des résultats fondée sur le Rappel, la Précision et le F_{score}

Toutes les exécutions du défi ont été évaluées en calculant le F_{score} (avec $\beta = 1$) :

$$F_{score}(\beta) = \frac{(\beta^2 + 1) \times Précision \times Rappel}{\beta^2 \times Précision + Rappel} \quad (3)$$

Sachant que la précision et le rappel sont fournies par les mesures suivantes :

$$Précision = \frac{\text{nb phrases correctement extraites}}{\text{nb phrases extraites}} \quad (4)$$

$$Rappel = \frac{\text{nb phrases correctement extraites}}{\text{nb phrases pertinentes}} \quad (5)$$

	tâche 1	**tâche 2**	**tâche 3**
équipe 1	0.870 *(1)*	0.884 *(1)*	0.880 *(1)*
équipe 2	0.860 *(2)*	0.852 *(2)*	0.866 *(2)*
équipe 3	0.820 *(3)*	0.821 *(3)*	0.819 *(3)*
équipe 4	0.760 *(4)*	0.742 *(6)*	0.746 *(6)*
équipe 5	0.751 *(5)*	0.755 *(5)*	0.755 *(5)*
équipe 6	0.732 *(6)*	0.794 *(4)*	0.788 *(4)*
équipe 7	0.562 *(7)*	0.559 *(8)*	0.573 *(7)*
équipe 8	0.494 *(8)*	0.521 *(9)*	0.507 *(9)*
équipe 9	0.493 *(9)*	0.560 *(7)*	0.563 *(8)*
équipe 10	0.325 *(10)*	0.307 *(10)*	0.305 *(11)*
équipe 11	0.177 *(11)*	0.177 *(11)*	0.417 *(10)*

TAB. 4 – Meilleurs F_{score} des différentes équipes pour chaque tâche. Le rang correspondant pour chaque tâche est indiqué entre parenthèses. Ces résultats sont triés par F_{score} décroissant sur la base de la première tâche.

Le F_{score} pour $\beta = 1$ se réécrit de la manière suivante :

$$F_{score} = \frac{2 \times \text{nb phrases correctement extraites}}{\text{nb phrases pertinentes} + \text{nb phrases extraites}} \quad (6)$$

Le tableau 4 et les figures 2, 3 et 4 présentent les meilleurs F_{score} obtenus par les différentes équipes pour chaque tâche. L'analyse des résultats détaillés (par exécution) sur la base du F_{score} avec $\beta = 1$ permet de voir que la plupart des équipes ont amélioré leurs résultats au fur et à mesure des tâches. Ainsi, l'ajout d'informations (noms de personnes, puis années) représente une aide réelle pour les différents systèmes représentés dans ce défi.

Le tableau 5 présente les meilleures mesures de précision et et de rappel pour chaque équipe. Par exemple, pour la première tâche, ce tableau montre que les équipes ayant un meilleur score de précision (équipe 5) et de rappel (équipe 9) n'ont pas obtenu les meilleurs scores en terme de F_{score} (voir tableau 4). Ceci s'explique par le fait que ces équipes ayant une précision (resp. rappel) de bonne qualité ont un rappel (resp. précision) décevant. Ainsi, pour avoir un F_{score} ayant une valeur importante, il est nécessaire d'avoir un bon compromis entre le rappel et la précision. Les deux équipes les mieux placées en terme de F_{score} sont les équipes 1 et 2 (voir tableau 4). Ainsi, ces dernières ont obtenu des valeurs élevées aussi bien en terme de précision que de rappel (voir tableau 5).

La figure 2 confirme que les résultats en terme de précision et de rappel sont singulièrement différents d'une équipe à l'autre qui peuvent avoir des F_{score} assez faibles. Par exemple, l'équipe 10 possède une précision très correcte alors que le rappel est faible et inversement le rappel de l'équipe 9 est très élevé et la précision est faible. Selon la tâche à effectuer, une équipe ayant une excellente précision (resp. rappel) peut être particulièrement recherchée. Ainsi, une manière différente d'évaluer le résultat global

	Précision		
	tâche 1	**tâche 2**	**tâche 3**
éq 5	0.941*(1)*	0.925 *(1)*	0.927 *(1)*
éq 1	0.883*(2)*	0.911 *(2)*	0.890 *(3)*
éq 2	0.880*(3)*	0.884 *(3)*	0.875 *(4)*
éq 3	0.868*(4)*	0.866 *(4)*	0.891 *(2)*
éq 4	0.801*(5)*	0.802 *(6)*	0.754 *(7)*
éq 10	0.778*(6)*	0.790 *(7)*	0.791 *(6)*
éq 6	0.666*(7)*	0.808 *(5)*	0.848 *(5)*
éq 9	0.532*(8)*	0.520 *(8)*	0.546 *(8)*
éq 7	0.413*(9)*	0.472 *(9)*	0.505 *(9)*
éq 8	0.365*(10)*	0.384 *(10)*	0.364 *(11)*
éq 11	0.226*(11)*	0.226 *(11)*	0.402 *(10)*

	Rappel		
	tâche 1	**tâche 2**	**tâche 3**
éq 9	0.915*(1)*	0.623 *(9)*	0.630 *(9)*
éq 7	0.900*(2)*	0.904 *(1)*	0.906 *(1)*
éq 1	0.858*(3)*	0.861 *(2)*	0.872 *(3)*
éq 2	0.856*(4)*	0.857 *(4)*	0.884 *(2)*
éq 4	0.849*(5)*	0.859 *(3)*	0.861 *(4)*
éq 6	0.812*(6)*	0.781 *(6)*	0.736 *(8)*
éq 8	0.783*(7)*	0.852 *(5)*	0.838 *(5)*
éq 3	0.777*(8)*	0.780 *(7)*	0.758 *(6)*
éq 5	0.755*(9)*	0.756 *(8)*	0.757 *(7)*
éq 10	0.207*(10)*	0.190 *(10)*	0.189 *(11)*
éq 11	0.145*(11)*	0.145 *(11)*	0.434 *(10)*

TAB. 5 – Meilleures mesures de précision et de rappel des différents équipes pour chaque tâche. Le rang correspondant pour chaque tâche est indiqué entre parenthèses.

des équipes consiste à utiliser le front de Pareto qui sera développé dans la section suivante.

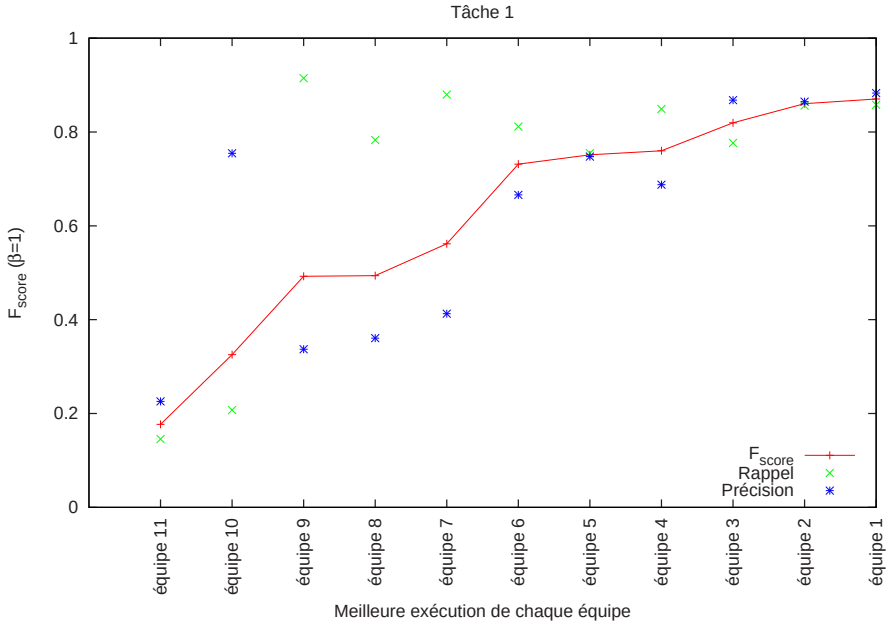

FIG. 2 – F_{scores} $(\beta = 1)$ pour les meilleures exécutions - Tâche 1.

4.2 Analyse des résultats fondée sur le front de Pareto

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 (ici précision et rappel). Les approches qui ne sont pas sur le front de Pareto sont dites "dominées".

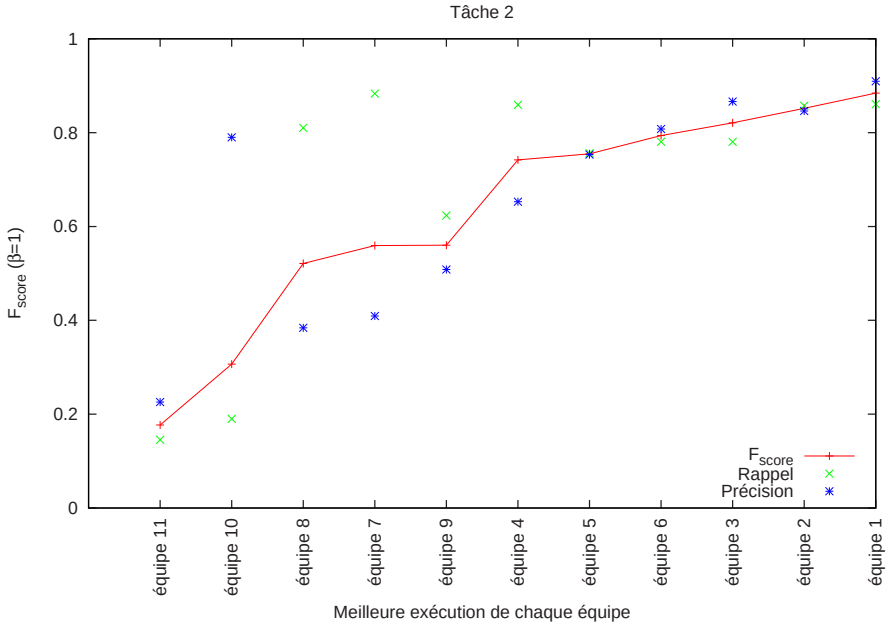

FIG. 3 – F_{scores} $(\beta = 1)$ pour les meilleures exécutions - Tâche 2.

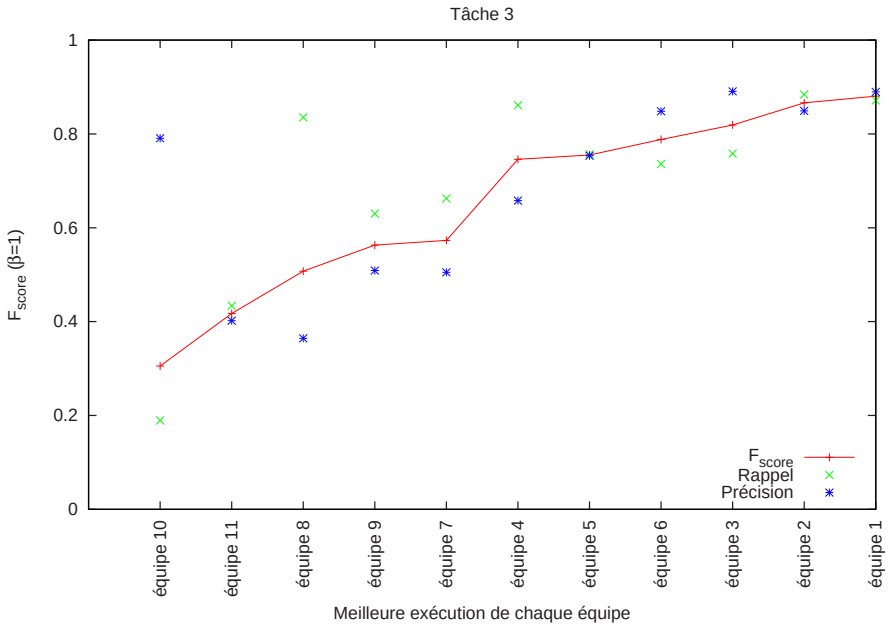

FIG. 4 – F_{scores} $(\beta = 1)$ pour les meilleures exécutions - Tâche 3.

Par exemple, deux des approches soumises par l'équipe numéro 5 pour la tâche 1 sont telles qu'elles présentent une précision optimale (par rapport aux autres résultats) pour un rappel faible (voir Figure 5). Mais il n'existe pas d'autres approches ayant à la fois une meilleure précision **et** un meilleur rappel.

L'analyse du front de Pareto associé à la tâche 1 (voir Figure 5) 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 F_{score} peut être modifié. Nous obtenons les mêmes résultats pour les tâches 2 et 3 (voir Figures 6 et 7).

Notons que l'équipe ayant le meilleur résultat en terme de F_{score} avec $\beta = 1$ est obtenu avec l'équipe se situant au plus près du point de coordonnées (1,1).

La courbe 5 présente l'ensemble des exécutions des équipes. Cette courbe montre que cinq équipes sont présentes sur le front de Pareto : les équipes 1, 2, 5, 7 et 9. Nous avons noté dans la section précédente que l'équipe 9 avait un rappel très élevé, ce qui explique la présence de cette équipe sur le front de Pareto.

Notons enfin que sur cette même courbe (Figure 5), les deux points relatifs à l'équipe 5 (évoqués précédemment) situés sur le front de Pareto ne représentent pas le F_{score} sélectionné pour le classement du défi. En effet, ces deux exécutions ont un F_{score} relativement faible (0.580 et 0.416 - voir tableau 7 de l'Annexe) comparativement à celui de la troisième exécution (0.751 - voir tableau 4 et tableau 7 de l'Annexe). Pourtant, ces deux exécutions appartiennent au front de Pareto, contrairement à la troisième qui présente la meilleure valeur de F_{score}.

Pour cette même équipe, une valeur de β de 0.2 permet d'obtenir un $F_{score} = 0.886$ pour l'approche présentant une précision de 0.926 et un rappel de 0.422. Cette valeur de β indique que la précision a un poids supérieur à celui du rappel. De plus, avec une telle valeur de β, cette équipe se retrouverait en première position pour le critère du F_{score}.

5 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

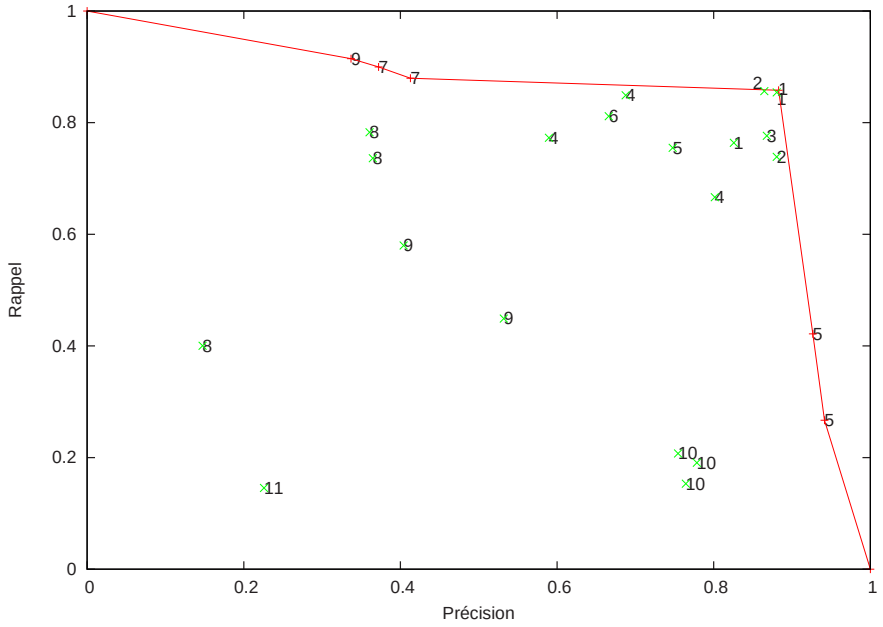

FIG. 5 – Front de Pareto pour la tâche 1.

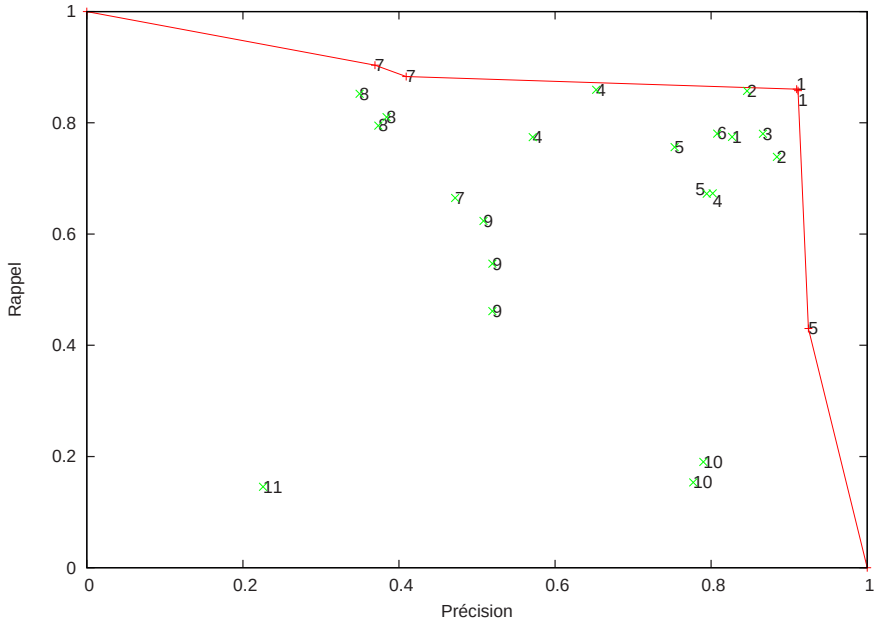

FIG. 6 – Front de Pareto pour la tâche 2.

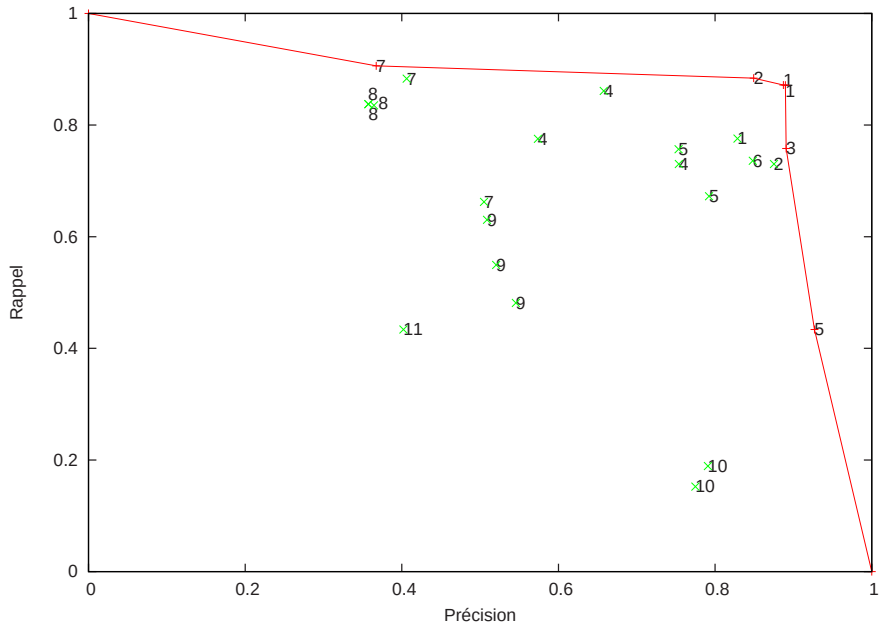

FIG. 7 – Front de Pareto pour la tâche 3.

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 en 2006. Comme l'a montré l'atelier consacré à DEFT'05 dans le cadre de la conférence TALN'05, DEFT'05 a réussi à fédérer la communauté francophone de fouille de textes. Lors de cet atelier, l'ensemble des participants a souhaité poursuivre l'initiative amorcée avec DEFT'05. Cependant, la majorité des chercheurs présents a souhaité que DEFT'06 mette plutôt l'accent sur des données réelles pour pouvoir ainsi mieux évaluer les différentes approches.

Les organisateurs de DEFT'05 se sont donc portés volontaires pour organiser la deuxième édition de DEFT. Trois corpus différents sont mis à la disposition des participants dans le cadre du défi DEFT'06. La nature du défi a également évolué puisqu'il s'agit de se focaliser sur la détection de segments thématiques clairement identifiés dans les corpus utilisés. L'identification de ces segments thématiques a été réalisée indépendamment de DEFT'06 soit par les auteurs des corpus, soit par la nature même des corpus ou enfin par des experts des domaines concernés. Cette expertise extérieure au défi permet de répondre à l'une des questions soulevées dans le cadre de DEFT'05 qui concernait la compétence du comité d'organisation dans l'analyse des discours politiques et leurs classements dans les catégories nationales ou internationales.

Références

[Amrani *et al.*, 2004] 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.

[Baayen *et al.*, 2002] Baayen, R. H., H. Van Halteren, A. Neijt, and F. Tweedie (2002), An experiment in authorship attribution. *In : Proceedings of JADT 2002. St. Malo*, pp. 29-37.

[Rudman, 1997] Joseph Rudman (1997), The State of Authorship Attribution Studies : Some Problems and Solutions, *Computers and the Humanities*, Volume 31, Issue 4, Jul 1997, pages 351-365.

[Rudolf, Świdziński, 2004] 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, 1993] Smadja F. (1993), Retrieving collocations from text : Xtract, *Computational Linguistics*, Vol. 191, p143-177.

[Soboroff, Harman, 2003] 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)*.

Summary

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 analysed in terms of their f_{score} and their place on the Pareto front.

Annexe

Numéro d'équipe	Laboratoire	Correspondant
1	LIA	**Marc Elbèze** Grégoire Moreau-de-Montcheuil Patrice Bellot Juan-Manuel Torres-Moreno
2	ENST	**Loïs Rigouste** Olivier Cappé François Yvon
3	LORIA UHP ESIAL	**Laurent Pierron** Coskun Durkal Sébastien Freydiger
4	LIA	**Alexandre Labadié** Yann Romero Laurianne Sitbon
5	CLIPS-Imag	**Loïc Maisonnasse** Caroline Tambellini
6	LIP6	**Frédéric Kerloch** Ludovic DENOYER Patrick Gallinari
7	LIMSI	**Nicolas Hernandez** Gabriel Illouz Benoit Habert
8	LGI2P	**Michel Plantié** Gérard Dray Alexandre Meimouni Pascal Poncelet Jacky Montmain
9	LIMSI	**Martine Hurault-Plantet** Michèle Jardino
10	LIRMM	**Jacques Chauché**
11	LaBRI	**Richard Moot** Patrick Henry Renaud Marley Maxime Amblard

TAB. 6 – Équipes ayant participées à DEFT'05 (les contacts de chaque équipe sont indiqués en gras).

Équipe	Exécution	*Précision*	*Rappel*	F_{score}
1	3	0.883	0.858	0.870
1	1	0.880	0.854	0.867
2	1	0.865	0.856	0.860
3	1	0.868	0.777	0.820
2	2	0.880	0.739	0.803
1	2	0.826	0.764	0.794
4	1	0.688	0.849	0.760
5	1	0.748	0.755	0.751
6	1	0.666	0.812	0.732
4	2	0.801	0.666	0.728
4	3	0.590	0.773	0.669
5	2	0.926	0.422	0.580
7	1	0.413	0.880	0.562
7	2	0.372	0.900	0.526
8	3	0.361	0.783	0.494
9	3	0.337	0.915	0.492
8	2	0.365	0.736	0.488
9	1	0.532	0.449	0.487
9	2	0.404	0.580	0.476
5	3	0.941	0.267	0.416
10	3	0.755	0.207	0.325
10	2	0.778	0.191	0.306
10	1	0.764	0.153	0.255
8	1	0.147	0.400	0.215
11	1	0.226	0.145	0.177

TAB. 7 – Résultats détaillés pour la tâche 1.

Équipe	Exécution	*Précision*	*Rappel*	F_{score}
1	1	0.909	0.861	0.884
1	3	0.911	0.858	0.884
2	1	0.846	0.857	0.852
3	1	0.866	0.780	0.821
2	2	0.884	0.739	0.805
1	2	0.827	0.775	0.800
6	1	0.808	0.781	0.794
5	1	0.753	0.756	0.755
4	1	0.653	0.859	0.742
4	2	0.802	0.673	0.732
5	3	0.794	0.672	0.728
4	3	0.571	0.774	0.657
5	2	0.925	0.430	0.587
9	1	0.508	0.623	0.560
7	2	0.409	0.883	0.559
7	3	0.472	0.665	0.552
9	2	0.520	0.547	0.533
7	1	0.369	0.904	0.524
8	3	0.384	0.810	0.521
8	1	0.374	0.795	0.508
8	2	0.350	0.852	0.496
9	3	0.520	0.461	0.489
10	2	0.790	0.190	0.307
10	1	0.777	0.153	0.256
11	1	0.226	0.145	0.177

TAB. 8 – Résultats détaillés pour la tâche 2.

Équipe	Exécution	*Précision*	*Rappel*	F_{score}
1	3	0.890	0.871	0.880
1	1	0.887	0.872	0.879
2	1	0.849	0.884	0.866
3	1	0.891	0.758	0.819
1	2	0.829	0.776	0.801
2	2	0.875	0.730	0.796
6	1	0.848	0.736	0.788
5	1	0.753	0.757	0.755
4	1	0.658	0.861	0.746
4	2	0.754	0.730	0.742
5	3	0.792	0.672	0.727
4	3	0.574	0.775	0.660
5	2	0.927	0.434	0.591
7	3	0.505	0.662	0.573
9	3	0.509	0.630	0.563
7	2	0.406	0.883	0.557
9	1	0.520	0.549	0.535
7	1	0.367	0.906	0.523
9	2	0.546	0.481	0.511
8	3	0.364	0.835	0.507
8	2	0.357	0.837	0.501
8	1	0.357	0.838	0.501
11	1	0.402	0.434	0.417
10	2	0.791	0.189	0.305
10	1	0.775	0.152	0.255

TAB. 9 – Résultats détaillés pour la tâche 3.

d1	d2	d3
i	e	i
n	p	e
e	i	
		r
s	r	p
p	m	n
L	L	m
t	n	s
m	t	L
r	é	t
o	u	o
J	J	J
i<>n	s	m<>i
e<>	m<>i	i<>n
o<>s	e<>	e<>
t<>r	r<>e	r<>e
<>J	p<>p	o<>s
i<>s	u<>p	t<>r
p<>i	t<>r	i<>s
s<>t	i<>s	s<>t
<>m	s<>t	<>m
L<>e	<>m	L<>e
J<>o	L<>e	J<>o
m<>i	J<>u	<>p
s<>p	i<>n	i<>e
r<>e	<>p	n<>i
n<>i	i<>e	p<>r
r<>e<>	n<>i	<>J
m<>i<>n	p<>r	p<>i
i<>s<>t	<>J	r<>
J<>o<>s	r<>	e<>m
n<>i<>s	e<>m	e<>r
<>m<>i	e<>r	s<>p
<>J<>o	p<>é	r<>e<>
e<> <>m	r<>e<>	<>p<>r
i<>n<>i	<>p<>r	m<>i<>n
e<> <>J	u<>p<>p	e<>r<>
t<>r<>e	m<>i<>n	i<>s<>t
s<>p<>i	e<>r<>	J<>o<>s
L<>e<>	i<>s<>t	e<>m<>i
s<>t<>r	e<>m<>i	n<>i<>s
o<>s<>p	n<>i<>s	<>m<>i
p<>i<>n	<>m<>i	<>J<>o
Jospin	m<>i<>e	m<>i<>e
ministre	r<>e<>m	r<>e<>m
ministre<>Jospin	i<>n<>i	i<>n<>i
Le<>ministre<>	e<> <>J	e<> <>J
Le<>ministre<>Jospin<>	i<>e<>r	i<>e<>r
	p<>p<>é	r<> <>m
	<>J<>u	t<>r<>e
	r<> <>m	s<>p<>i
	t<>r<>e	e<> <>p
	e<> <>p	L<>e<>
	L<>e<>	s<>t<>r
	J<>u<>p	p<>i<>n
	s<>t<>r	o<>s<>p
	p<>r<>e	p<>r<>e
	Juppé	Jospin
	premier	premier
	ministre	ministre
	Le<>premier	Le<>premier
	premier<>ministre	ministre<>Jospin
	ministre<>Juppé	premier<>ministre
	Le<>premier<>ministre	premier<>ministre<>Jospin
	premier<>ministre<>Juppé	Le<>premier<>ministre

TAB. 10 – Détail des n-grammes de l'exemple donné en section 2.3.1, page 6, le symbole <> correspond au séparateur des éléments des n-grammes.