

Présentation et résultats du défi fouille de texte DEFT2010 Où et quand un article de presse a-t-il été écrit ?

Cyril Grouin¹ Dominic Forest² Lyne Da Sylva²
Patrick Paroubek¹ Pierre Zweigenbaum¹

(1) LIMSI-CNRS, BP133, 91403 Orsay Cedex, France

(2) École de Bibliothéconomie et des sciences de l'information, Université de Montréal,
C.P. 6128, succursale Centre-ville, Montréal H3C 3J7, Canada

cyril.grouin@limsi.fr, dominic.forest@umontreal.ca, lyne.da.sylva@umontreal.ca,
patrick.paroubek@limsi.fr, pierre.zweigenbaum@limsi.fr

Résumé. Cet article détaille l'édition 2010 du défi fouille de texte. Deux tâches ont été proposées : identifier la décennie de publication d'un extrait d'article de presse paru entre 1800 et 1944, et identifier le pays puis le titre du journal de parution d'un article de presse. Les résultats sont faibles et éparés pour la première tâche (meilleure F-mesure de 0,338 pour une moyenne de 0,193) témoignant de la difficulté à traiter ce type de données. Les résultats de la seconde tâche sont corrects pour l'identification du pays (meilleure F-mesure de 0,932 pour une moyenne de 0,767) et moyens pour l'identification du titre du journal (meilleure F-mesure de 0,741 pour une moyenne de 0,489). Les résultats démontrent que les systèmes classent aisément des documents propres sur une échelle restreinte de valeurs ; en revanche, ces systèmes appellent des améliorations pour traiter des documents bruités.

Abstract. This paper describes the DEFT 2010 text mining challenge. Two tasks have been presented : to identify the publication decade of a press article extract published from 1800 to 1944, and to identify the country and the newspaper name of a press article. Results are low and scattered for the first task (0.338 best F-measure and 0.193 mean F-measure) showing difficulty to process this kind of data. Results of the second task are corrects when identifying the country (0.932 best F-measure and 0.767 mean F-measure) while they are medium in identifying the newspaper name (0.741 best F-measure and 0.489 mean F-measure). The results show that the systems easily classify clean documents on restricted scale. Nevertheless, these systems need to be improved to process noisy documents.

Mots-clés : Campagne d'évaluation, classification automatique, internationalisation, variation linguistique, diachronie, diatopie.

Keywords: Evaluation campaign, automatic classification, internationalization, linguistic variation, diachrony, diatopy.

1 Introduction

L'édition 2010 du défi fouille de texte (DEFT) est la sixième de cette campagne annuelle francophone. Pour la première fois, la campagne a été co-organisée par deux équipes de deux pays, l'une française (le LIMSI à Orsay¹), l'autre québécoise (l'EBSI à Montréal²). Cette particularité nous a incité à nous orienter vers la variation linguistique diachronique et diatopique du français.

La principale contrainte dans l'organisation d'une campagne sur un tel sujet demeure l'obtention de corpus illustrant ces différents phénomènes linguistiques. Deux catégories de corpus – a priori aisément disponibles – nous ont paru intéressantes pour ce type d'étude : les textes de lois et les articles de presse. Dans les faits, seuls les corpus de presse se sont révélés accessibles pour deux pays : la France et le Canada (Québec). Cette édition se focalise donc sur deux axes d'étude linguistique d'un corpus : la variation en diachronie et la variation selon l'origine géographique. Deux tâches ont ainsi été définies et proposées :

- L'identification de la décennie de publication d'un extrait d'article de presse sur une période d'un siècle et demi (de 1800 à 1944) parmi cinq journaux français ;
- L'identification du pays de parution d'un article puis du journal d'où provient l'article étudié.

Développer des méthodes permettant de dater des documents sur une large échelle telles que les décennies constitue un préalable pour l'étude de variations linguistiques en diachronie sur des documents non datés. Alors que les outils actuels permettent de traiter efficacement des données propres, disposer d'outils permettant de travailler sur des données bruitées, résultant d'un système de reconnaissance des caractères par exemple, représente la prochaine étape à franchir. (Galibert *et al.*, 2010) ont ainsi développé un système de reconnaissance des entités nommées sur des articles de journaux OCRisés dans le cadre des évaluations Quero et soulignent les besoins d'adapter les méthodes à ce type de données. L'un des domaines d'application de l'identification du pays d'origine d'un texte, et par extension l'identification de l'auteur d'un document, concerne au niveau juridique l'anonymat dans des textes, soit du point de vue de la levée de l'anonymat, soit au contraire pour s'assurer du maintien de l'anonymat d'un auteur (et de l'impossibilité de remonter à l'auteur d'un document). La détection des particularités linguistiques propres à un pays permet de mieux gérer l'internationalisation d'applications du traitement automatique des langues.

Dans cet article, nous reviendrons sur le déroulement de cette édition (section 2), puis nous présenterons pour chaque tâche, la constitution des corpus et les résultats obtenus par les participants (section 3).

2 Déroulement du défi

2.1 Calendrier

Les inscriptions ont été ouvertes le 25 janvier 2010 après parution d'appels à participation sur les principales listes de diffusion du traitement automatique des langues. L'accès aux données d'entraînement a été rendu possible à partir du 31 mars pour les équipes ayant signé et renvoyé les licences d'utilisation des corpus. Lors des précédentes éditions, la période de test courait sur deux semaines, une fenêtre de trois jours devant être choisie dans cet intervalle pour faire tourner les systèmes sur les données de test. Les par-

¹LIMSI : Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, UPR3251 du CNRS.

²EBSI : École de bibliothéconomie et des sciences de l'information, Université de Montréal (Québec).

Participants attendant généralement les derniers jours pour profiter d'une phase d'apprentissage la plus longue possible, nous avons restreint la période de test à la semaine du 31 mai au 4 juin. Les participants ont ensuite eu deux semaines à compter de la réception de leurs résultats pour rédiger l'article présentant les méthodes et ressources utilisées. Les résultats individuels de l'évaluation ont par ailleurs été communiqués aux participants quelques jours après la soumission des fichiers produits par les systèmes, la présentation des résultats globaux étant réservée pour l'atelier de clôture.

2.2 Participations

À l'instar des éditions 2005 et 2008, cette campagne s'est déroulée dans le cadre de la conférence TALN. Dix équipes ont fait acte de candidatures, six ont accédé aux corpus et soumis des résultats :

- CLAC *Computational Linguistics at Concordia* (S. Mokhov),
- CLUL *Centro de Linguística da Universidade de Lisboa* (M. Génereux),
- LIA *Laboratoire d'Informatique d'Avignon* (S. Oger, M. Rouvier, N. Camelin, R. Kessler, F. Lefèvre, J. M. Torres-Moreno),
- LIMSI *Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur* (P. Albert, F. Bardin, M. Delorme, N. Devos, S. Papazoglou, J. Simard),
- LINA *Laboratoire d'Informatique de Nantes Atlantique* (L. Monceaux et A. Tartier),
- LUTIN *Laboratoire des Usages en Technologies d'Information Numérique* (A. El Ghali et Y. V. Hoareau).

Une précision s'impose quant à la participation du LIMSI (laboratoire co-organisateur de la campagne) : plusieurs étudiants du laboratoire ont souhaité participer. Ils ont en conséquence strictement été tenus à l'écart de l'organisation de la campagne et n'ont bénéficié d'aucun traitement de faveur par rapport aux autres participants.

2.3 Mesures d'évaluation des résultats

Les deux tâches peuvent être envisagées comme relevant d'une classification dans laquelle les éléments à classer sont :

- pour la tâche 1, un extrait de journal parmi quinze classes (une classe correspondant à la décennie d'appartenance de l'extrait) : 1800, 1810, 1820, 1830, 1840, 1850, 1860, 1870, 1880, 1890, 1900, 1910, 1920, 1930 et 1940 ; la décennie 1800 couvre ainsi les années 1800 à 1809, etc. ;
- pour la tâche 2, un article de journal parmi deux classes (correspondant aux pays d'origine) : France vs. Québec, et deux sous-classes par pays (correspondant aux noms des journaux dans lequel l'article a paru) : *L'Est Républicain* et *Le Monde* pour la France, *Le Devoir* et *La Presse* pour le Québec.

Chaque fichier de résultat pour une tâche a été évalué en calculant la F-mesure sur toutes les classes de cette tâche avec $\beta = 1$, ce qui ne privilégie ni la précision ni le rappel, mais un équilibre entre les deux.

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

La précision et le rappel sur les classes d'une tâche sont ici calculés suivant la macro-moyenne (Nakache & Métais, 2005) dans laquelle chaque classe compte à égalité avec les autres, qu'elle ait un fort ou un faible effectif. Lors de la constitution des corpus, nous avons cependant veillé à équilibrer les classes des différents corpus.

F-mesure pondérée Dans la F-mesure classique, une seule classe peut être attribuée à chaque document. Cependant, un système de classification automatique peut attribuer à un document une distribution de probabilité sur les différentes classes au lieu de lui attribuer une seule classe. L'indice de confiance est cette probabilité pour un document d'appartenir à une catégorie donnée. Dans la F-mesure pondérée, la précision et le rappel pour chaque classe sont pondérés par l'indice de confiance. Ce qui donne :

$$\text{Précision}_i = \frac{\sum_{\text{attribué correct. } i=1}^{\text{Nombre attribué correct. } i} \text{indice de confiance}_{\text{attribué correct. } i}}{\sum_{\text{attribué } i=1}^{\text{Nombre attribué } i} \text{indice de confiance}_{\text{attribué } i}}$$

$$\text{Rappel}_i = \frac{\sum_{\text{attribué correct. } i=1}^{\text{Nombre attribué correct. } i} \text{indice de confiance}_{\text{attribué correct. } i}}{\text{nombre de documents appartenant à la classe } i}$$

Avec :

- Nombre attribué correct._{*i*} : nombre de documents attribués correct._{*i*} appartenant effectivement à la classe *i* et auxquels le système a attribué un indice de confiance non nul pour cette classe ;
- Nombre attribué_{*i*} : nombre de documents attribués_{*i*} auxquels le système a attribué un indice de confiance non nul pour la classe *i*.

La F-mesure pondérée est ensuite calculée à l'aide des formules de la F-mesure classique.

Macro-moyenne

$$\text{Précision} = \frac{\sum_{i=1}^n \left(\frac{TP_i}{(TP_i + FP_i)} \right)}{n} \quad \text{Rappel} = \frac{\sum_{i=1}^n \left(\frac{TP_i}{(TP_i + FN_i)} \right)}{n}$$

Avec :

- TP_i = nombre de documents correctement attribués à la classe *i* ;
- FP_i = nombre de documents faussement attribués à la classe *i* ;
- FN_i = nombre de documents appartenant à la classe *i* et non retrouvés par le système ;
- n = nombre de classes.

3 Présentation détaillée des tâches

Les corpus rassemblés pour les deux tâches concernent des articles de presse. Alors que la tâche 1 se focalise sur des articles de presse ancienne, la tâche 2 repose sur des articles de presse contemporaine.

3.1 Tâche 1. Identification de la décennie

3.1.1 Constitution des données

Présentation Depuis plusieurs années, la Bibliothèque Nationale de France a entrepris une démarche de numérisation de son fond documentaire. Le résultat de cette numérisation est librement accessible depuis le portail Gallica³. Dans le domaine de la presse ancienne, une vingtaine de titres français est disponible sur la période 1800–1944 au format image (PDF ou JPG). Une reconnaissance de caractères a été appliquée pour cinq titres seulement : *Le Journal des Débats* (1800–1805), *Le Journal de l'Empire* (1805–1814), *Le Journal des Débats politiques et littéraires* (1814–1944), *Le Figaro* (1826–1942), et *La Croix* (1880–1944). Les trois premiers titres se succèdent dans le temps et correspondent au même journal sous différents noms (en parallèle des changements politiques dans le pays). Le résultat de cette reconnaissance est proposé dans des fichiers textuels et dans les fichiers PDF multi-couche. Nous avons constitué notre corpus sur la base des versions textuelles de ces cinq titres.

Chaîne de traitements Dans un premier temps, nous avons rapatrié sur nos serveurs l'intégralité des archives textuelles de ces cinq journaux (un fichier par page numérisée, cf. tableau 1).

Journal	J. Débats	J. Empire	J. Débats pol et litt	Le Figaro	La Croix
Fichiers	7 060	11 777	175 313	95 944	48 407

FIG. 1 – Nombre total de fichiers textuels rapatriés par journal

Chaque fichier a ensuite fait l'objet d'une segmentation en portions de 300 mots (dans le sens d'une suite de caractères comprise entre deux espaces). Cette taille a été définie à l'issue des évaluations humaines – réalisées sur des portions de 1100 à 1400 mots (se reporter section 3.1.2) – que nous avons jugées trop volumineuses et nécessaires de réduire.

Puisque les fichiers en notre possession correspondent au résultat d'une reconnaissance de caractères, nous ne disposons d'aucun indice de début et de fin d'article. En conséquence, les segments produits peuvent intégrer aussi bien un seul extrait d'un long article tronqué, que plusieurs petits articles s'enchaînant (un ensemble de brèves par exemple). Ces segments peuvent également être interrompus au milieu d'une phrase. En revanche, nous avons rétabli les césures de manière à réduire le nombre de mots coupés.

Deux types de segments ont été éliminés. D'abord les segments contenant des caractères inutilisés en français et faussement identifiés par la reconnaissance de caractères (le tilde ~, l'accent circonflexe sans voyelle ^, l'esperluette & et l'astérisque *) puis les segments contenant plus de vingt chiffres ; ces derniers correspondent généralement aux résultats de la bourse ou aux programmes du théâtre intégrant heures et adresses. Enfin, les années facilement reconnaissables (« 1857 » mais ni « !8b2 », ni « !92i ») ont été remplacées par une balise <annee /> (voir figure A.1 pour un exemple de document). Nous donnons dans les tableaux 2 et 3 la répartition des segments par journal et décennie.

À l'issue de cette phase de préparation, ces segments de journaux constituent les documents types que les participants du défi ont eu à classer.

³Portail Gallica : <http://gallica.bnf.fr/>, site visité le 17 mai 2010.

	1800	1810	1820	1830	1840	1850	1860	1870
J. Débats	4145							
J. Empire	725	654						
J. Débats pol et litt		1739	4102	22767	29661	62723	61976	40293
Le Figaro			2				40	139

FIG. 2 – Nombre de segments de 300 mots par journal et par décennie (de 1800 à 1870)

	1880	1890	1900	1910	1920	1930	1940
J. Débats pol et litt	34035	33692	43029	29039	33579	29973	8440
Le Figaro	25	5766	15420	38994	57874	78933	13556
La Croix	5112	679	10578	3800	18030	40742	14682

FIG. 3 – Nombre de segments de 300 mots par journal et par décennie (de 1880 à 1940)

Pour chaque journal et chaque décennie, un tirage aléatoire a ensuite été réalisé pour répartir les documents entre données d’entraînement et données de test. Un seuil maximal de 421 documents par décennie a cependant été défini pour deux raisons. En premier lieu, pour équilibrer le nombre de documents par décennie et éviter ainsi toute sur ou sous-représentation. Notons toutefois qu’à l’intérieur d’une décennie, nous n’avons pas contrôlé les années qui ont été extraites (voir figures 4 et 5 pour la ventilation des documents par année et par corpus). En second lieu, ce seuil nous a permis de disposer de corpus finaux de taille raisonnable, soit un total de 6315 documents répartis entre apprentissage (3594 documents) et test (2721 documents) selon le ratio habituel de 60% des données pour l’apprentissage et 40% pour le test. Rapporté au niveau des décennies, cela représente 252 documents par décennie pour le corpus d’apprentissage et 169 documents pour le corpus de test.

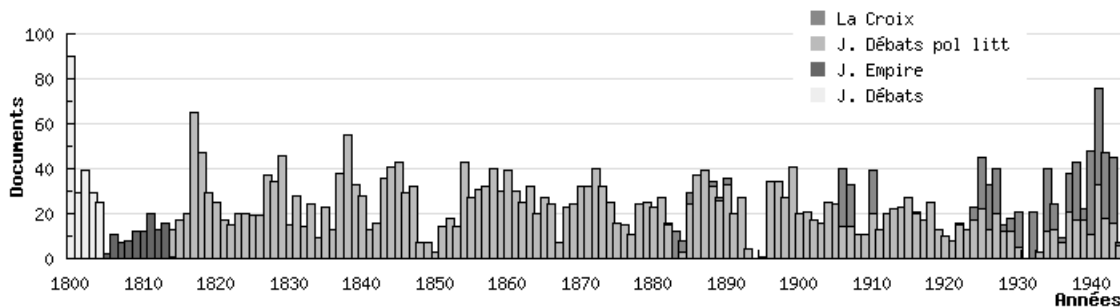


FIG. 4 – Tâche 1 : nombre de documents par année et par journal (données d’entraînement)

Afin d’éprouver la robustesse des systèmes des participants, nous avons fait le choix de réserver les articles d’un journal (*Le Figaro*) pour le corpus de test, les quatre autres journaux étant disponibles à la fois dans les corpus d’entraînement et de test. Les participants ont été informés de ce dispositif sur le site Internet de la campagne sans que ne soit précisé le nom du journal réservé pour le corpus de test. Les graphiques 4 et 5 détaillent le nombre de documents par décennie et par journal pour les corpus d’apprentissage et de test.

PRÉSENTATION ET RÉSULTATS DU DÉFI FOUILLE DE TEXTE DEFT2010

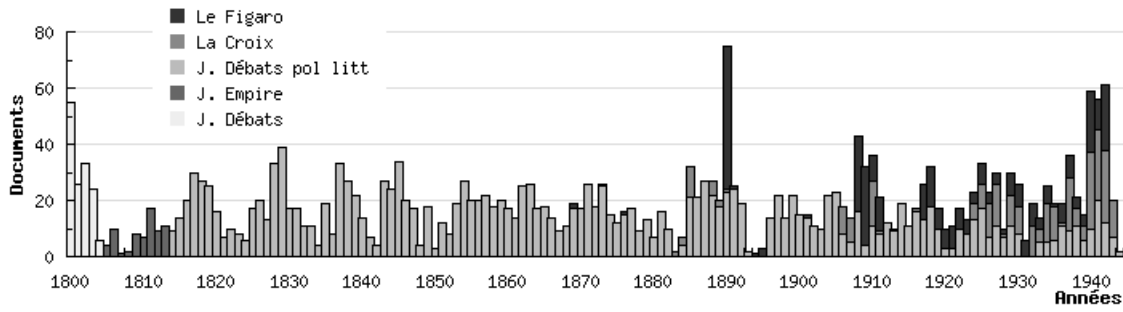


FIG. 5 – Tâche 1 : nombre de documents par année et par journal (données de test)

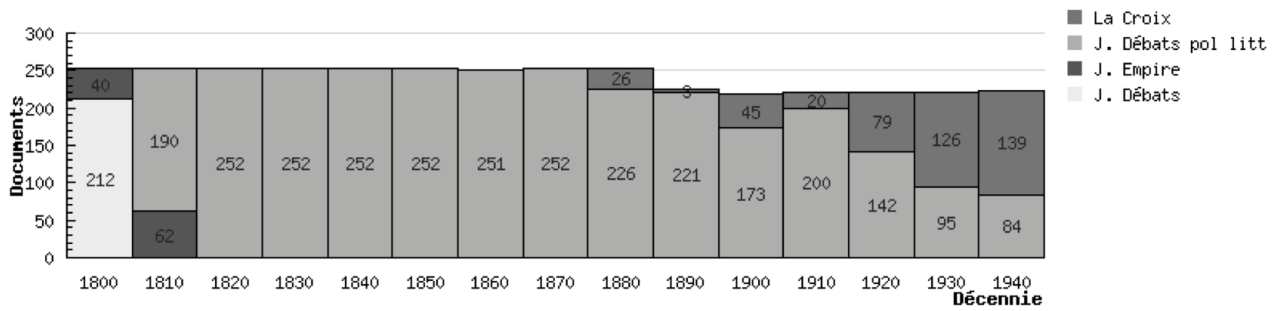


FIG. 6 – Tâche 1 : nombre de documents par décennie et par journal (données d'entraînement)

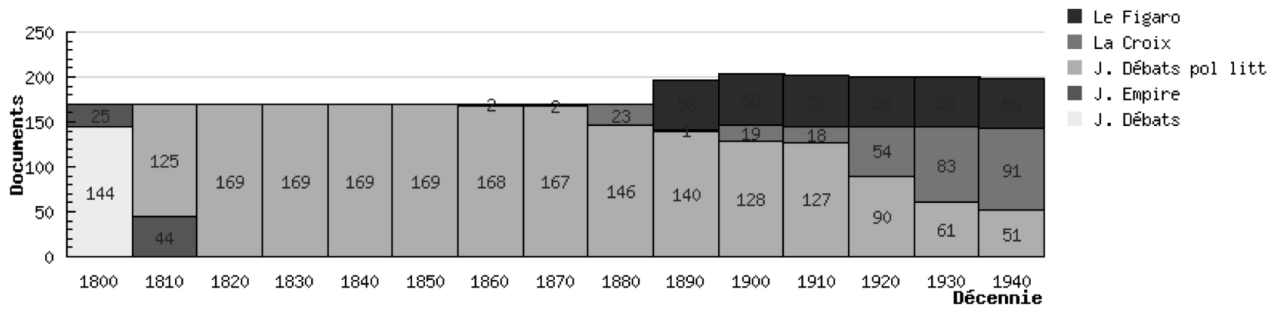


FIG. 7 – Tâche 1 : nombre de documents par décennie et par journal (données de test)

Précisons toutefois que, par suite d'une erreur de programmation, le fait de réserver les articles d'un journal pour le seul corpus de test a engendré un déséquilibre mineur dans le nombre de documents réellement disponibles par décennie, comme l'attestent les graphiques 6 et 7. À partir de 1890 (décennie correspondant à une disponibilité élevée d'articles du *Figaro*, cf. tableaux 2 et 3), le nombre de documents par décennie décroît dans le corpus d'apprentissage (passant de 252 à 221 documents) tandis qu'il croît dans le corpus de test (passant de 169 à 198 documents). Cette observation ayant été effectuée après distribution du corpus d'apprentissage aux participants et ne portant pas trop à conséquence, nous avons maintenu ce déséquilibre.

3.1.2 Évaluation humaine

Une petite évaluation humaine a été réalisée comme suit : six segments de 200 lignes provenant de six éditions du *Figaro* ont été proposés à plusieurs juges humains qui ont reçu pour consigne d'identifier la décennie de publication parmi quatre décennies possibles (1910, 1920, 1930 et 1940). Les résultats obtenus par ces juges ont varié de 0,125 à 0,875 en terme de F-mesure globale tandis que les coefficients Kappa ont varié de -0,15 à 0,78 entre juge et référence et de 0,08 à 0,33 entre juges (soit un meilleur accord entre juges qu'entre chaque juge et la référence). La lecture de ces chiffres doit s'accompagner de la plus grande prudence en raison du trop faible nombre de documents que les juges humains ont eu à classer. Il nous est cependant apparu que des portions de 200 lignes (entre 1100 et 1400 mots par portion) étaient trop volumineuses ; en tant qu'évaluateurs humains, nous arrivions à extraire des indices pour identifier la décennies dans la première partie de chaque document. La taille a donc été revue à la baisse lors de la préparation des corpus.

3.1.3 Résultats

Les résultats des participants sur cette première tâche varient de 0,053 à 0,338 de F-mesure sur les meilleures soumissions (lignes grisées du tableau 8) avec des disparités assez fortes entre participants. Avec une F-mesure moyenne de 0,193 et une F-mesure médiane de 0,181, les résultats témoignent de la difficulté de la tâche proposée et s'expliquent en partie par le bruit inhérent à l'OCRisation des données d'origine (voir figure A.1).

3.2 Tâche 2. Identification de l'origine géographique

La seconde tâche du défi se compose de deux pistes complémentaires : identifier le pays de parution d'un article, puis le journal dans lequel l'article a paru.

3.2.1 Constitution des données

Nous avons rassemblé des corpus d'articles de journaux provenant de deux pays, les articles étant issus de deux titres différents par pays. Le corpus québécois se compose d'articles provenant des journaux *La Presse* et *Le Devoir* ; ils ont été obtenus auprès de l'agence CEDROM-SNi⁴. Le corpus français intègre

⁴CEDROM-SNi : <http://www.cedrom-sni.com/>, visité le 19 mai 2010.

PRÉSENTATION ET RÉSULTATS DU DÉFI FOUILLE DE TEXTE DEFT2010

Participant	Soumission	Macro rappel	Macro précision	Macro F-mesure	Rang
CLUL	1	0,171	0,198	0,183	3
CLUL	2	0,169	0,190	0,179	
CLUL	3	0,163	0,188	0,174	
CLAC	1	0,116	0,107	0,111	5
LINA	1	0,051	0,050	0,050	
LINA	2	0,053	0,052	0,053	6
LINA	3	0,053	0,052	0,053	
LIA	1	0,293	0,295	0,294	2
LIA	2	0,258	0,260	0,259	
LIA	3	0,266	0,264	0,265	
Lutin	1	0,108	0,126	0,116	
Lutin	2	0,157	0,155	0,156	
Lutin	3	0,178	0,182	0,180	4
LIMSI	1	0,299	0,297	0,298	
LIMSI	2	0,340	0,336	0,338	1
LIMSI	3	0,313	0,308	0,310	

FIG. 8 – Résultats obtenus par les participants sur la tâche 1. La meilleure soumission est sur fond grisé.

des articles du *Monde* fournis par l'agence ELDA⁵ et de *L'Est Républicain* fournis par le CNRTL⁶. Les corpus de ces quatre journaux couvrent les années 1999, 2002 et 2003. Nous avons par ailleurs restreint les sujets traités à deux domaines thématiques : les informations générales (politique nationale et internationale) et les articles de sports. Ces domaines thématiques ont été identifiés dans les corpus au moyen des informations présentes dans les méta-données⁷. Le choix de ces deux domaines repose sur l'hypothèse selon laquelle les articles de sport seraient davantage identifiables géographiquement que les articles de politique, par exemple en étudiant les disciplines sportives couvertes (le baseball et le hockey au Québec, le football et le rugby en France). Pour chaque journal, nous avons limité le nombre d'articles à 750 par domaine thématique (soit un maximum de 1500 articles par journal sur les trois années couvertes). Nous avons essayé de conserver un équilibre dans les corpus finaux entre pays (France 46,5% des articles vs. Québec 53,5%), entre journaux (*L'Est Républicain* 22,3% des articles, *La Presse* 27,3%, *Le Devoir* 26,2% et *Le Monde* 24,2%), et entre catégories thématiques (Informations générales 51,1% des articles vs. Sports 48,9%) sans recourir à un égalitarisme absolu.

Peu de traitements préparatoires a été appliqué sur ces corpus à l'exception d'un traitement typographique et d'une sélection d'articles du *Monde*. Pour les articles dont les premiers mots étaient imprimés en capitales d'imprimerie, nous avons rétabli ces mots en minuscules en essayant de conserver les capitales pour les acronymes. Le corpus du *Monde* étant le plus fourni des quatre journaux, nous avons fixé un seuil de

⁵Evaluations and Language resources Distribution Agency (ELDA) : <http://www.elda.org/>, visité le 19 mai 2010.

⁶Centre National de Ressources Textuelles et Lexicales (CNRTL) : <http://www.cnrtl.fr/>, visité le 19 mai 2010.

⁷Les secteurs de rédaction d'informations générales « ING » et de sports nationaux « SNA » pour *L'Est Républicain*, les secteurs de rédaction internationale « INT » et de sports « SPO » pour *Le Monde*. Pour le corpus *La Presse*, nous avons rassemblé les rubriques « actualités », « arts et culture », « autres » et « politique nationale et internationale » dans la catégorie « Informations générales » et les rubriques « société et tendance » et « sports et loisirs » dans la catégorie « Sports ». Enfin, pour le corpus du *Devoir*, nous avons rassemblé les rubriques « Actualités », « La Une » et « Politique nationale et internationale » dans la catégorie « Informations générales » tandis que la rubrique « Sports et loisirs » a été versée dans la catégorie « Sports ».

300 caractères minimum pour conserver un article. Aucune anonymisation n'a été produite sur les corpus de cette tâche.

3.2.2 Résultats

Présentation générale. Cinq équipes ont participé à la seconde tâche. Les résultats sur l'identification du pays (deux classes) sont supérieurs à ceux obtenus pour l'identification du titre du journal (quatre classes). Sur les meilleures soumissions de ces équipes (lignes grisées du tableau 9), la F-mesure moyenne est de 0,767 et la médiane de 0,792 pour la piste d'identification du pays, tandis que la F-mesure moyenne est de 0,489 et la médiane de 0,462 pour la piste d'identification du journal.

Participant	Soumission	Piste	Macro rappel	Macro précision	Macro F-mesure	Rang
CLAC	1	Pays	0,532	0,532	0,532	5
		Journaux	0,278	0,143	0,189	
CLAC	2	Pays	0,532	0,532	0,532	
		Journaux	0,278	0,143	0,189	
CLUL	1	Pays	0,854	0,861	0,858	2
		Journaux	0,607	0,655	0,630	
CLUL	2	Pays	0,845	0,853	0,849	
		Journaux	0,611	0,653	0,631	
CLUL	3	Pays	0,845	0,852	0,849	
		Journaux	0,598	0,648	0,622	
LIA	1	Pays	0,933	0,931	0,932	1
		Journaux	0,742	0,739	0,741	
LIA	2	Pays	0,820	0,821	0,820	
		Journaux	0,379	0,380	0,379	
LIA	3	Pays	0,964	0,965	0,964	
		Journaux	0,708	0,702	0,705	
LINA	1	Pays	0,721	0,725	0,723	4
		Journaux	0,419	0,430	0,425	
LINA	2	Pays	0,692	0,695	0,694	
		Journaux	0,396	0,413	0,404	
LINA	3	Pays	0,685	0,688	0,687	
		Journaux	0,393	0,414	0,403	
Lutin	1	Pays	0,749	0,775	0,762	
		Journaux	0,419	0,429	0,424	
Lutin	2	Pays	0,796	0,800	0,798	
		Journaux	0,447	0,445	0,446	
Lutin	3	Pays	0,793	0,791	0,792	3
		Pays	0,458	0,466	0,462	

FIG. 9 – Résultats obtenus par les participants sur la tâche 2. La meilleure soumission est sur fond grisé.

Des disciplines sportives géographiquement marquées ? Lors de la préparation du corpus, nous avons émis l'hypothèse que les articles relatifs à quatre disciplines sportives autoriseraient une identification plus aisée du pays (section 3.2.1). Afin de vérifier cette hypothèse, nous avons procédé à une évaluation portant uniquement sur les 1216 documents du corpus de test émergeant dans la catégorie « Sports ». Ces documents se répartissent comme suit en termes de discipline sportive par pays (voir figure 10).

Pays	Baseball	Football	Hockey	Rugby	Autres
France	0	127	5	41	380
Québec	53	36	112	2	460

FIG. 10 – Répartition des articles sportifs par discipline et par pays sur le corpus de test de la tâche 2.

La meilleure soumission de chaque participant a fait l'objet d'une évaluation sur deux jeux de données : d'une part sur les 376 articles sportifs relevant des quatre disciplines pré-identifiées, et d'autre part sur les 840 autres articles sportifs ne traitant pas de ces quatre disciplines. Sur les cinq participants (figure 11), quatre obtiennent des résultats légèrement supérieurs sur le jeu de données des articles traitant des quatre disciplines sportives identifiées. Notre hypothèse de départ est donc vérifiée.

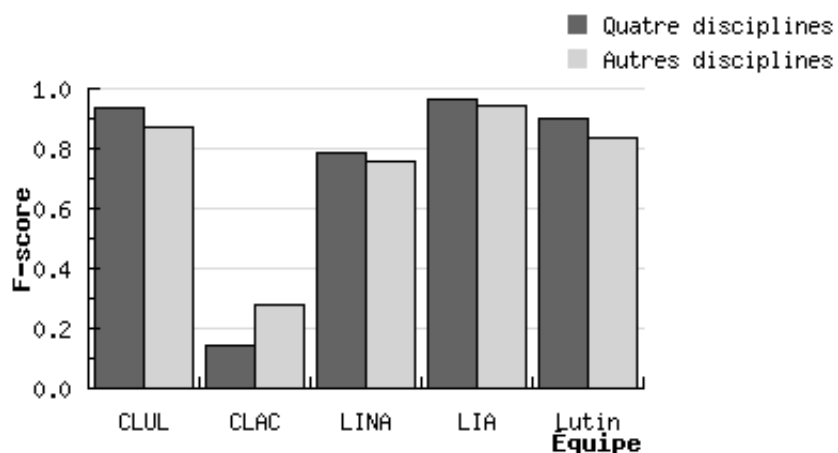


FIG. 11 – Tâche 2 : identification des pays

3.3 Méthodes des participants

L'ensemble des participants a considéré les deux tâches proposées comme relevant d'une classification de documents, dans quinze classes de décennies pour la première tâche, dans deux classes de pays et quatre classes de journaux pour la seconde tâche.

Chaque équipe a généralement mobilisé des approches statistiques, soit de manière exclusive, soit en les combinant avec des approches symboliques.

Parmi les hypothèses suivies, certains ont mis en évidence les termes saillants par décennie et les croisances et décroissances de termes dans le temps (Généreux, 2010). D'autres ont fusionné les résultats de

différentes techniques (Oger *et al.*, 2010) : repérage d'entités nommées, apprentissage par SVM et modèles de langues, et validations croisées. La combinaison de plusieurs catégories par regroupement de classes attendues a également fait l'objet d'une approche (El Ghali & Hoareau, 2010). Des techniques d'analyse issues de l'oral ont été adaptées et testées à l'écrit, la chaîne MARF par (Mokhov, 2010), un système de reconnaissance d'entités nommées adapté aux données bruitées des retranscriptions de l'oral (Oger *et al.*, 2010).

Des ressources externes ont parfois été produites, telles que des lexiques de termes spécifiques aux catégories « sports » et « informations générales » (Généreux, 2010) ou des listes d'entités nommées de type événement (Monceaux & Tartier, 2010).

Au niveau linguistique, une étude fine des réformes de l'orthographe (formations de l'imparfait et du pluriel) a été réalisée par (Albert *et al.*, 2010) conduisant à la mise en place de filtres sur les années de ces réformes. Des essais de corrections manuelles et automatiques (par règles et par correcteur orthographiques) ont également été réalisés (El Ghali & Hoareau, 2010).

4 Conclusion

Deux tâches de classification de documents ont été proposées aux participants de cette nouvelle édition du défi fouille de texte. La première, diachronique, concernait l'identification de la décennie de publication d'extraits de journaux OCRisés parus entre 1800 et 1944. La seconde, diatopique, visait l'identification du pays et du journal dans lequel a pu un article complet.

La première tâche s'est révélée difficile (F-mesure moyenne de 0,193 et médiane de 0,181), combinant à la fois un nombre élevé de classes (quinze) et une qualité moyenne des documents proposés (lié à la reconnaissance des caractères). Les approches linguistiques (tentatives de correction orthographique et étude des réformes de l'orthographe dans le temps) ont permis d'améliorer les résultats obtenus. Notons que l'utilisation de techniques de traitement des retranscriptions de la parole ont été adaptées aux besoins de la tâche.

La seconde tâche, portant sur des données de meilleure qualité et pour un nombre de classes plus réduit (deux pour l'identification du pays, quatre pour celle du journal) a mieux été réussie (F-mesures moyennes de 0,767 et de 0,489 respectivement pour chaque piste, et F-mesures médianes de 0,792 et 0,462). L'utilisation de lexiques thématiques a de nouveau permis l'amélioration des résultats.

Remerciements

Nous exprimons nos remerciements les plus sincères aux agences et institutions ayant mis à disposition les corpus utilisés pour cette édition du défi fouille de texte : la Bibliothèque Nationale de France au travers de son portail Gallica pour la tâche d'identification des décennies, les agences Cedrom-SNi pour les corpus presse québécois, ELDA pour le corpus du Monde et le CNRTL pour le corpus de l'Est Républicain.

L'organisation de cet atelier a bénéficié du soutien financier du projet DoXa (projet CapDigital convention DGE n° 08 2 93 0888). Nous remercions les organisateurs de TALN pour l'organisation matérielle. Enfin, nous remercions les participants pour les approches et les idées originales qu'ils ont pu mettre en œuvre.

Références

- ALBERT P., BADIN F., DELORME M., DEVOS N., PAPAZOGLU S. & SIMARD J. (2010). Décennie d'un article de journal par analyse statistique et lexicale. In *Actes de TALN2010*, Montréal. Atelier DEFT2010.
- EL GHALI A. & HOAREAU Y. V. (2010). μ -Alida : expérimentations autour de la catégorisation multi-classes basée sur Alida. In *Actes de TALN2010*, Montréal. Atelier DEFT2010.
- GALIBERT O., QUINTARD L., ROSSET S., ZWEIGENBAUM P., NÉDELLEC C., AUBIN S., GILLARD L., RAYSZ J.-P., POIS D., TANNIER X., DELÉGER L. & LAURENT D. (2010). Named and Specific Entity Detection in Varied Data : The Quæro Named Entity Baseline Evaluation. In N. CALZOLARI, K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODJIK, S. PIPERIDIS, M. ROSNER & D. TAPIAS, Eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).
- GÉNÉREUX M. (2010). Classification de textes en comparant les fréquences lexicales. In *Actes de TALN2010*, Montréal. Atelier DEFT2010.
- MOKHOV S. (2010). A MARF Approach to DEFT 2010 : L'Approche MARF à DEFT 2010. In *Actes de TALN2010*, Montréal. Atelier DEFT2010.
- MONCEAUX L. & TARTIER A. (2010). Utilisation d'outils linguistiques pour trouver la date ou l'origine d'un fragment textuel. In *Actes de TALN2010*, Montréal. Atelier DEFT2010.
- NAKACHE D. & MÉTAIS E. (2005). Evaluation : nouvelle approche avec juges. In *INFORSID*, p. 555–570, Grenoble.
- OGER S., ROUVIER M., CAMELIN N., KESSLER R., LEFÈVRE F. & TORRES-MORENO J.-M. (2010). Système du LIA pour la campagne DEFT'10 : datation et localisation d'articles de presse francophones. In *Actes de TALN2010*, Montréal. Atelier DEFT2010.

A Corpus

A.1 Tâche 1

Nous donnons ci-après un exemple de document issu du corpus d'apprentissage sur la tâche 1 d'identification des décennies de parution. La classe de référence associée au document figure entre balises <periode>. Les années aisément identifiables ont été remplacées par une balise <annee />. Aucun traitement n'a été appliqué pour réduire le bruit lié à l'OCRisation.

Dans le corpus de test, aucune méta-information n'est disponible : ni la date de publication (sic !), ni le nom du journal d'où provient l'extrait (puisque les cinq titres utilisés ne sont disponibles que sur une partie de la période étudiée).

```
<portion id="891">
  <meta>
    <journal>Le Journal des Débats politiques et littéraires</journal>
    <date annee="1927" mois="03" jour="31" />
  </meta>
  <periode>1920</periode>
  <texte>
    de deuxième classe; La création à Paris d'un office international du vin; La création d'un corps d'ingénieurs de l'aéronautique et d'un corps d'ingénieurs adjoints et d'agents techniques de l'aéronautique; Des modifications à la loi du 31 décembre 1913 sur les monuments historiques. Séance demain jeudi. Le groupe de la Gauche démocratique, réuni hier sous la présidence de M. Bienvenu Martin, a émis à l'unanimité le vœu que les conseils généraux, lors de leur prochaine session, soient invités à se prononcer en faveur du rétablissement du scrutin uninominal. Sur une intervention de M. Labrousse, un débat auquel ont pris part MM. Fernand Rabier, Pierre Marraud, Labrousse, Machet, a été institué sur la question du vote des femmes actuellement pendante devant une des commissions du Sénat. Il résulte de ce débat que presque tous les membres présents se sont montrés hostiles à la réforme. Le groupe a délibéré aussi au sujet des intentions que l'on a prêtées à M. Albert Sarraut, ministre de l'intérieur; M. Albert Sarraut aurait déclaré à une délégation de l'Union pour le suffrage des femmes que, l'un des adversaires les plus résolus du vote féminin en 1913, il considérait actuellement la situation comme toute différente et qu'il comprenait fort bien qu'au point de vue de la justice, comme au point de vue de la défense de leurs intérêts et des réformes sociales, les femmes aient le droit de voter. Aussi le groupe a-t-il décidé de recueillir l'opinion du ministre de l'intérieur sur cette question.
  </texte>
</portion>
```

A.2 Tâche 2

Voici un document extrait du corpus d'apprentissage de la tâche 2 portant sur l'identification du pays et du titre du journal. Dans le corpus de test, seule la catégorie thématique de parution (« Sports » vs. « Informations générales ») demeure.

```
<article id="489">
  <meta>
    <journal>L'Est Républicain</journal>
    <pays>France</pays>
    <categorie>Informations générales</categorie>
  </meta>
  <titre>La diplomatie française se remobilise</titre>
  <texte>Le ministre français des Affaires Etrangères, Dominique de
  Villepin va se lancer, dimanche, dans une mission difficile au
  Proche-Orient qui vit un des pire moments de son histoire, et où la
  diplomatie française a traditionnellement du mal à s&apos;affirmer
  face au poids américain. M. de Villepin aura des consultations au
  Caire avec le président Moubarak, avant de se rendre en Israël et
  dans les territoires palestiniens, puis en Arabie saoudite. Il ne
  manquera pas de sujets difficiles et contradictoires lors de ses
  entretiens avec ses interlocuteurs. Plus que le désir d&apos;obtenir
  un Etat, les Palestiniens veulent avant tout mettre fin à
  l&apos;occupation. Une demande totalement rejetée par Israël qui
  accentue sa mainmise sur la Cisjordanie sur fond d&apos;attentats
  -suicides. La visite de M. de Villepin est la première d&apos;un
  ministre français des Affaires étrangères dans la région depuis
  septembre 2001. Pour cause d&apos;élections, la France est restée
  ces derniers temps en retrait sur le dossier israélo-palestinien
  qui mobilise les capitales arabes et occidentales. A la veille de
  sa visite, Dominique de Villepin a exprimé toute son "horreur" et
  sa "révolte", après les derniers attentats, estimant que "le peuple
  palestinien ne devait pas être l&apos;otage des terroristes". Dans
  ce contexte, diplomates et analystes soulignent l&apos;absence
  totale de perspective politique. En outre, face à la méfiance
  d&apos;Israël, la marge de manœuvre de la France et de l&apos;Union
  européenne est pour le moins limitée.</texte>
</article>
```