DEFT2010

Actes du sixième DÉfi Fouille de Textes

Proceedings of the Sixth DEFT Workshop

23 juillet 2010

Montréal, Canada

DEFT2010

Actes du sixième DÉfi Fouille de Textes

23 juillet 2010
Montréal, Canada

Préface

DEFT2010, sixième édition de la campagne d'évaluation en fouille de textes, portera sur les variations diachroniques et géographiques en corpus de presse francophones. L'atelier de clôture se tiendra à Montréal dans le cadre de la conférence TALN 2010.

Un locuteur francophone natif est capable de détecter dans une conversation des expressions spécifiques à un pays (par exemple au niveau des nombres septante et nonante en Belgique et en Suisse contre soixante-dix et quatre-vingt-dix en France et au Québec, et huitante en Suisse vs. quatre-vingts dans les trois autres pays).

Un lecteur est également capable de mobiliser des connaissances linguistiques, culturelles et historiques pour identifier la période (sur une échelle plus ou moins grande) de parution d'un article (en identifiant un événement particulier et/ou des tournures linguistiques ou des entités nommées jugées représentatives d'une période donnée).

Comme tout acte de communication, les documents ont une origine et un public visé ; leur nature, c'est-à-dire leurs contenu, niveaux de langue, etc. en dépend fortement. Dans cette édition du défi fouille de textes, nous nous intéressons à l'origine des documents, plus particulièrement à l'époque et au lieu de leur création.

Dans ce cadre, nous proposons plusieurs pistes distinctes et indépendantes.

Piste 1.

Cette piste, relative à la variation diachronique, concerne l'identification de la décennie de publication d'extraits d'articles français d'une taille de 300 mots. Les extraits de ce corpus couvrent une période comprise entre 1800 et 1944.

Le corpus d'apprentissage se composera d'extraits (300 mots) d'articles de quatre titres de journaux différents, le corpus de test intègrera des extraits provenant de ces quatre mêmes titres plus un cinquième titre absent du corpus d'apprentissage, de manière à éprouver la robustesse des systèmes.

Piste 2.

L'identification de l'origine géographique de chaque document (pays d'origine) constituera la seconde piste de cette campagne. Elle reposera sur des corpus de presse rassemblant plusieurs titres provenant de France et du Québec.

Présentation générale

Pour ces deux pistes, les participants ont eu la possibilité d'utiliser des ressources externes (linguistiques, historiques, etc.) qu'ils doivent obligatoirement déclarer. En ce qui concerne plus spécifiquement la piste 1, les ressources provenant de Gallica n'ont pas été autorisées.

Les participants ont été invités à participer aux deux pistes. Il est cependant possible de ne participer qu'à une seule des pistes.

Des corpus d'apprentissage ont été fournis aux participants inscrits, à partir du 31 mars 2010. Ces corpus sont composés de 60% des corpus d'origine. Les 40% de corpus restants ont été utilisés pour le test. Le test s'est déroulé sur une semaine, du 31 mai au 4 juin. Les participants ont bénéficié de trois jours pour appliquer, sur les corpus de test, les méthodes mises au point sur les corpus d'apprentissage et nous retourner leurs résultats d'analyse.

Cyril Grouin et Dominic Forest, *co-présidents du Comité de Programme*

Partenaires

Fournisseurs de corpus

Agence Cedrom-SNi, *corpus La Presse, Le Devoir*

CNRTL, *corpus L'Est Républicain*

ELDA, *corpus Le Monde*

Bibliothèque Nationale de France, portail Gallica, *corpus d'archives de presse*

Comités

Présidents

Cyril Grouin (LIMSI–CNRS, Orsay)

Dominic Forest (EBSI, Université de Montréal)

Comité de programme

Catherine Berrut (LIG, Grenoble)

Guillaume Cleuziou (LIFO, Orléans)

Lyne Da Sylva (EBSI, Université de Montréal)

Guy Denhière (EPHE, Paris)

Marc El Bèze (LIA, Avignon)

Dominic Forest (EBSI, Université de Montréal)

Patrick Gallinari (LIP6, Paris)

Cyril Grouin (LIMSI–CNRS, Orsay)

Thierry Hamon (LIPN, Villetaneuse)

Fidelia Ibekwe-SanJuan (Université Lyon III, Lyon)

Patrick Paroubek (LIMSI–CNRS, Orsay)

Pascal Poncelet (LIRMM, Montpellier)

Christian Roche (LISTIC, Annecy)

Mathieu Roche (LIRMM, Montpellier)

Pascale Sébillot (IRISA, Rennes)

François Yvon (LIMSI–CNRS, Orsay)

Pierre Zweigenbaum (LIMSI–CNRS, Orsay)

Comité d'organisation

Cyril Grouin (LIMSI–CNRS, Orsay)

Dominic Forest (EBSI, Université de Montréal)

Lyne Da Sylva (EBSI, Université de Montréal)

Table des matières

Préface iii
Partenaires v
Comités vii
Table des matières ix
Programme xi

Présentation et résultats **1**

Présentation et résultats du défi fouille de texte DEFT2010. Où et quand un article de presse a-t-il été écrit ? *Cyril Grouin, Dominic Forest, Lyne Da Sylva, Patrick Paroubek et Pierre Zweigenbaum* 3

Méthodes des participants **19**

Utilisation d'outils linguistiques pour trouver la date ou l'origine d'un fragment textuel. *Laura Monceaux et Annie Tartier* 21

A MARF Approach to DEFT 2010. *Serguei Mokhov* 35

μ-Alida : expérimentations autour de la catégorisation multi-classes basée sur Alida. *Adil El Ghali et Yann Vigile Hoareau* 51

Classification de textes en comparant les fréquences lexicales. *Michel Généreux* 57

Système du LIA pour la campagne DEFT'10 : datation et localisation d'articles de presse francophones. *Stanislas Oger, Mickael Rouvier, Nathalie Camelin, Rémy Kessler, Fabrice Lefèvre et Juan-Manuel Torres-Moreno* 69

Décennie d'un article de journal par analyse statistique et lexicale. *Pierre Albert, Flora Badin, Maxime Delorme, Nadège Devos, Sophie Papazoglou et Jean Simard* 85

Index **97**

Index des auteurs 99
Index des mots-clés 101

Programme

Vendredi 23 juillet 2010

9.30–10.00 *Accueil des participants*

SESSION I — PRÉSENTATION ET RÉSULTATS

10.00–10.30 Présentation et résultats du défi fouille de texte DEFT2010. Où et quand un article de presse a-t-il été écrit ? *Cyril Grouin, Dominic Forest, Lyne Da Sylva, Patrick Paroubek et Pierre Zweigenbaum*

SESSION II — MÉTHODES DES PARTICIPANTS

10.30–11.00 Utilisation d'outils linguistiques pour trouver la date ou l'origine d'un fragment textuel. *Laura Monceaux et Annie Tartier*

11.00–11.30 *Pause*

11.30–12.00 A MARF Approach to DEFT 2010. *Serguei Mokhov*

12.00–12.30 μ-Alida : expérimentations autour de la catégorisation multi-classes basée sur Alida. *Adil El Ghali et Yann Vigile Hoareau*

12.30–14.00 *Repas*

14.00–14.15 *Présentation de Cedrom-SNi*

14.15–14.45 Classification de textes en comparant les fréquences lexicales. *Michel Généreux*

14.45–15.15 Système du LIA pour la campagne DEFT'10 : datation et localisation d'articles de presse francophones. *Stanislas Oger, Mickael Rouvier, Nathalie Camelin, Rémy Kessler, Fabrice Lefèvre et Juan-Manuel Torres-Moreno*

15.15–15.45 *Pause*

15.45–16.15 Décennie d'un article de journal par analyse statistique et lexicale. *Pierre Albert, Flora Badin, Maxime Delorme, Nadège Devos, Sophie Papazoglou et Jean Simard*

SESSION III — DISCUSSION ET CONCLUSION

16.15–17.15 Discussion sur l'édition 2010 et pistes pour l'édition 2011.

17.15 *Clôture de l'atelier*

Présentation et résultats

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

Cyril Grouin[1] Dominic Forest[2] Lyne Da Sylva[2]
Patrick Paroubek[1] Pierre Zweigenbaum[1]
(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 épars 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 Orsay[1]), l'autre québécoise (l'EBSI à Montréal[2]). 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 Quæro 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 précédents 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-

[1]LIMSI : Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, UPR3251 du CNRS.
[2]EBSI : École de bibliothéconomie et des sciences de l'information, Université de Montréal (Québec).

ticipants attendant généralement les dernier 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. Généreux),
- 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. Badin, 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.

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

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

$$\text{Précision} = \frac{\sum_{i=1}^{n}\left(\frac{TPi}{(TPi+FPi)}\right)}{n} \quad \text{Rappel} = \frac{\sum_{i=1}^{n}\left(\frac{TPi}{(TPi+FNi)}\right)}{n}$$

Avec :
- TPi = nombre de documents correctement attribués à la classe i ;
- FPi = nombre de documents faussement attribués à la classe i ;
- FNi = 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 Gallica[3]. 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.

[3]Portail 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.

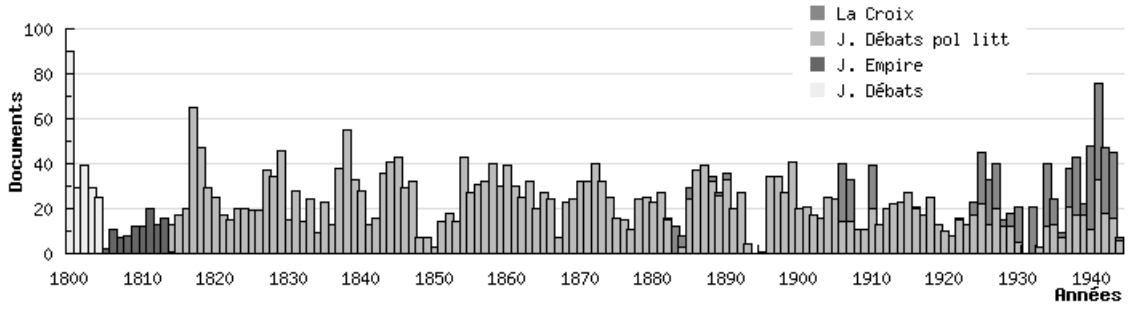


FIG. 4 – Tâche 1 : nombre de documents par année et par journal (données d'entraînement)

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.

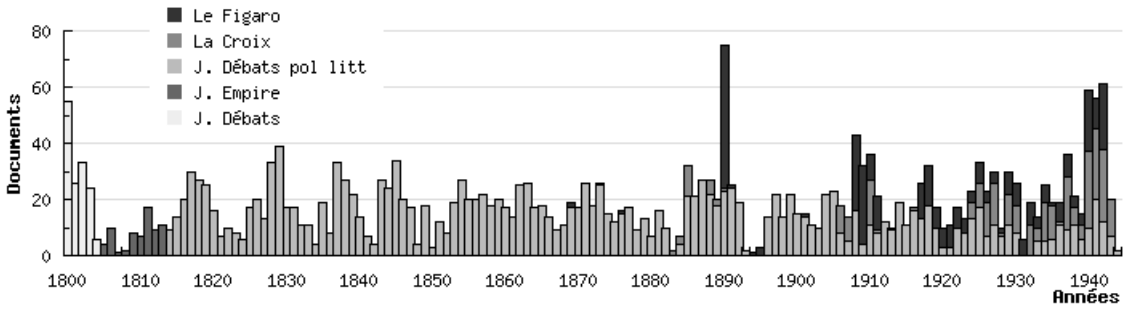


FIG. 5 – Tâche 1 : nombre de documents par année et par journal (données de test)

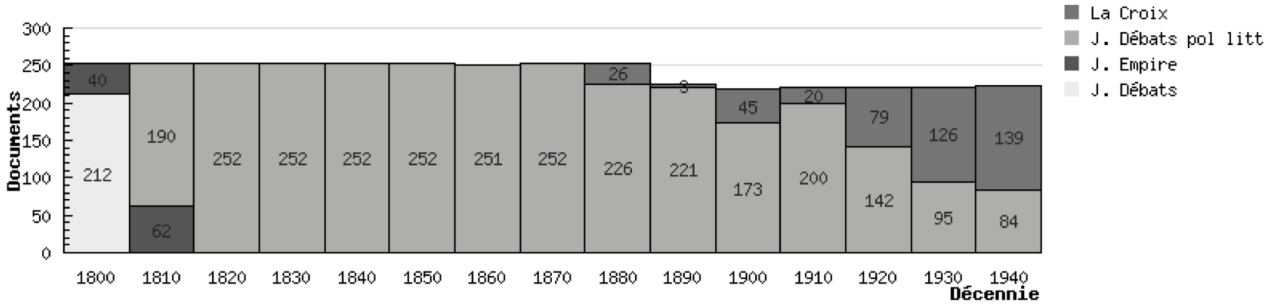


FIG. 6 – Tâche 1 : nombre de documents par décennie et par journal (données d'entraînement)

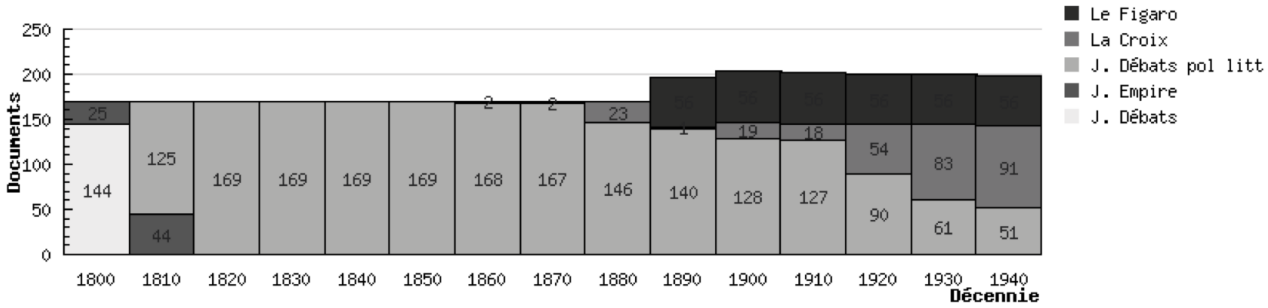


FIG. 7 – Tâche 1 : nombre de documents par décennie et par journal (données de test)

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-SNi[4]. Le corpus français intègre

[4]CEDROM-SNi : `http://www.cedrom-sni.com/`, visité le 19 mai 2010.

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 ELDA[5] et de *L'Est Républicain* fournis par le CNRTL[6]. 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-données[7]. 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

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

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

[7]Les secteurs de rédaction d'informations générales « ING » et de sports nationaux « SNA » pour *L'Est Républicain*, les secteurs de rédaction international « 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 équipe 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 avions 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 émargeant 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.

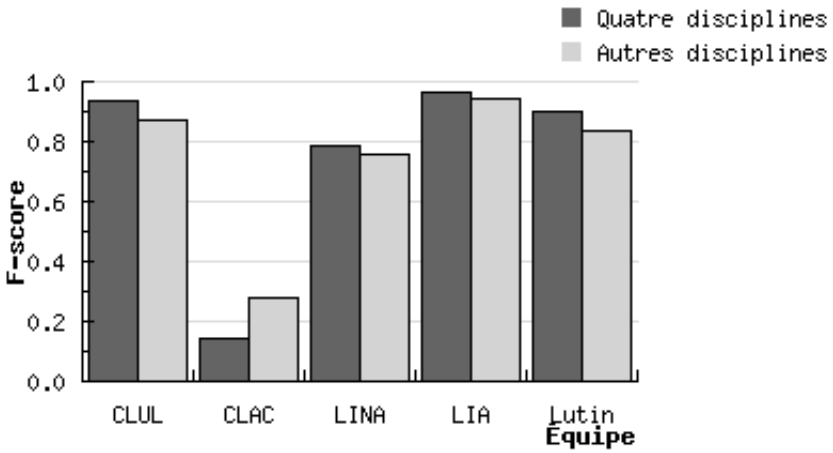


FIG. 11 – Tâche 2 : identification des pays

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 croissances 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 restranscriptions 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., PAPAZOGLOU 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 internatio
  nal du vin;  La création d'un corps d'ingénieurs de  l'aéronautique
  et d'un corps d'ingénieurs ad joints et d'agents techniques de
  l'aéronau tique;  Des modifications à la loi du 3! décembre  !9:3
  sur les monuments historiques.  Séance demain jeudi.  Le groupe de
  !a .Gauche démocratique,  réuni hier sous la présidence de M.
  Bienvenu Martin, a émis à l'unanimité le vo:u que les  conseils
  généraux, lors de leur prochaine ses sion, 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 a M.
  Albert  Sarraut, ministre de l'intérieur; M. Albert  Sarraut aurait
  déclaré à une délégation de I' &lt;: Uunion pour le suffrage des
  femmes x  que, l'un des adversaires les plus résolus du  vote féminin
  en <annee />. il considérait actuelle ment la situation comme toute
  difïérentê  et qu'il comprenait fort bien qu'au point de  vue de ta
  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-i) 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>
```

Méthodes des participants

Utilisation d'outils linguistiques pour trouver la date ou l'origine d'un fragment textuel

Laura Monceaux Annie Tartier
LINA, UMR 6241, Université de Nantes, 2 rue de la Houssinière, 44322 Nantes Cedex 03, France
laura.monceaux@univ-nantes.fr, annie.tartier@univ-nantes.fr

Résumé. Cet article décrit notre approche des deux tâches proposées, puis la mise en place de méthodes et stratégies pour affecter automatiquement une décennie ou une origine à un fragment textuel. Les méthodes appliquées sont basées sur le repérage d'éléments linguistiques tels que les entités nommées et les termes complexes. La stratégie a été choisie à partir de tests réalisés sur des extraits des corpus d'apprentissage. L'article se termine par une analyse des résultats obtenus et une ouverture vers des pistes plus prometteuses.

Abstract. This paper decribes our approch of the two proposed tasks, our methods and strategies to find automatically a decade or an origin for a piece of text. The methods that we applied are founded on marking linguistic elements such as named entities and complex terms. The choice of our strategy arises from tests made upon training corpora extracts. The paper ends with an analysis of the results we obtained, and with an somes propositions of improvement.)

Mots-clés : corpus d'apprentissage, corpus d'évaluation, extraction d'entités.

Keywords: training corpus, evaluation corpus, entities extraction.

1 Introduction

Cet article, plutôt technique, a pour objectif de présenter les méthodes que nous avons mises en œuvre pour répondre aux deux tâches de fouille de texte qui nous étaient proposées. Une première section présente notre approche de ces tâches et l'élaboration d'une méthode d'apprentissage. La section suivante est la description technique des phases d'apprentissage qui utilisent des sorties d'outils linguistiques. Elle est suivie d'une section qui explique le traitement des corpus d'évaluation . Les deux dernières sections présentent la stratégie adoptée pour générer les fichiers de sortie attendus. Nous terminons sur une tentative d'évaluation de nos résultats, au vu des éléments qui nous ont été renvoyés. Pour ne pas alourdir l'article, et parce que nous avons appliqué des méthodes similaires, nous avons pris le parti de ne pas séparer systématiquement la présentation des deux tâches. Cependant, lorsque c'est nécessaire, nous précisons ce qui est spécifique à l'une ou à l'autre.

2 Approche et élaboration des méthodes d'apprentissage

Les deux tâches proposées ont consisté à assigner une *classe* à chaque *portion* ou *article* d'un corpus d'évaluation. Pour chaque tâche un corpus d'apprentissage et un corpus d'évaluation, de structures analogues, nous ont été fournis.

2.1 Tâche n^o 1 : datation des portions

La première tâche est une tentative de datation automatique des *portions*. En effet les classes à affecter sont les quinze décennies qui constituent la période 1800-1944. Il faut noter que les *portions* des corpus sont des fragments textuels et non des articles à part entière. Ils peuvent commencer ou se terminer par une phrase incomplète et balayer plusieurs articles courts qui se suivent dans l'édition d'origine. Il faut noter aussi une forte dégradation du texte, due au procédé de numérisation, erreurs sur certains caractères et mots coupés à mauvais escient.

Les éléments d'un article de presse qui évoquent sa date de production sont sans doute d'abord les événements qui y sont mentionnés. C'est pourquoi, en plus du corpus d'apprentissage qui nous a été fourni, nous avons cherché ce qui se rapprochait le plus d'une base d'événements et nous nous sommes tournées vers les ressources historiques publiées par l'encyclopédie Wikipedia. Il existe en effet des pages qui listent les événements ayant eu lieu chaque année[1].

La question suivante concerne le repérage des événements. Les marqueurs les plus probables sont les *entités nommées*. Certains *termes complexes* comme, par exemple, "abolition de l'esclavage" peuvent aussi être de bons marqueurs parce qu'ils évoquent quelque chose de spécifique qui peut caractériser un événement. Nous avons donc décidé de caractériser les *portions* du corpus en extrayant ces deux catégories d'unités lexicales. Par crainte d'avoir une "couverture" trop faible des *portions* avec seulement les *entités nommées* et *termes complexes*, nous y avons ajouté les noms communs et les verbes présents dans les *portions*.

[1] Par exemple, la page de l'année 1827 se trouve à l'url `http://fr.wikipedia.org/wiki/1827`

2.2 Tâche n° 2 : affectation d'une origine à un article

Dans la seconde tâche il faut affecter à des *articles* deux informations dépendantes l'une de l'autre. Il s'agit d'une part du nom du journal dont est extrait l'article, d'autre part du pays d'origine de ce journal. Il est clair que si le nom du journal est connu avec certitude, il détermine celui du pays. Mais, une réflexion a priori peut laisser penser qu'il est plus facile de détecter les spécificités de la langue de chaque pays. Après avoir parcouru un certain nombre d'articles du corpus d'apprentissage qui nous a été fourni, nous n'avons pas repéré de spécificités notoires de la langue française de France ou de celle du Québec. Nous avons donc fait le choix de ne travailler que sur le nom du journal et d'en déduire automatiquement le nom du pays. Nous avons donc travaillé avec quatre classes *classes* qui sont les quatre noms de journaux. N'ayant pas non plus observé de différence notoire entre entre les langues de chaque pays pour les articles sportifs et pour ceux concernant les informations générales, nous n'avons pas utilisé la catégorie des articles qui nous était fournie dans les corpus.

Le corpus d'apprentissage proposé est constitué d'articles complets, récents, munis d'un titre. Pour classer un article "dans un journal", nous avons cherché à repérer dans quel journal du corpus d'apprentissage son lexique était le plus présent. Se contenter des mots simples aurait sans doute été peu discriminant, c'est pourquoi, en plus des noms communs et des verbes, nous avons considéré les *termes complexes* et les *entités nommées*. Nous n'avons pas distingué le titre du corps de l'article considérant que les éléments d'un titre se retrouvent en général dans le corps de l'article.

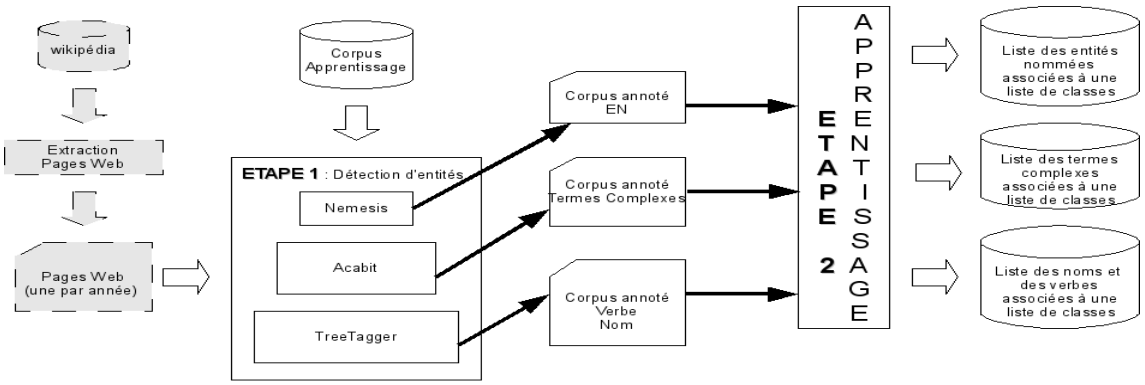


FIG. 1 – Apprentissage d'entités par classe

2.3 Plan de travail

D'un point de vue technique nous avons donc conduit un travail d'apprentissage similaire pour les deux tâches. Il se résume selon le plan ci-dessous et est détaillé dans la section suivante :

- Corpus d'apprentissage :
  - corpus fournis par DEFT2010,
  - pour la tâche n° 1 uniquement : corpus constitué par les pages d'événements de chaque année publiées dans *Wikipedia*.
- Entités extraites pour représenter une portion :
  - entités nommées,
  - termes complexes,
  - noms communs, et verbes.

2.4 Les outils

Pour réaliser les repérages d'entités dans les corpus nous avons utilisé les trois outils suivants dont les deux premiers ont été conçus au LINA :

– Les entités nommées ont été extaites avec NEMESIS, logiciel de reconnaissance, identification et catégorisation automatiques des entités nommées du français (Fourour, 2002).
– Les termes complexes ont été extraits avec ACABIT, programme d'acquisition de terminologie prenant en entrée un texte annoté linguistiquement et retournant une liste ordonnée de candidats termes (Daille, 2003). La vocation première de ce logiciel est de s'appliquer à des corpus de spécialité. Nous en avons un peu détourné l'usage en l'appliquant à un corpus journalistique, au risque, bien sûr, de dégrader les résultats.
– Les noms, verbes, adjectifs et adverbes ont été repérés par le logiciel TREETAGGER, lemmatiseur et outil d'annotation d'un texte en différentes parties du discours (Schmid, 1994).

Les documents résultant de l'apprentissage ont été générés avec des programmes, écrits spécialement pour cetta campagne, en Java, Perl ou Xslt.

3 Mise en œuvre de l'apprentissage

Dans cette section, et dans les suivantes, nous utilisons le terme générique *entité* pour désigner, selon le cas, une *entité nommée*, un *terme complexe*, un *nom commun* ou un *verbe*.

3.1 Construction d'une ressource externe pour la tâche n° 1

Nous avons téléchargé les 145 pages de Wikipédia correspondant chacune aux événements d'une année entre 1800 et 1944, frontières temporelles de l'étude. Puis nous avons fusionné ces pages par décennies, de manière à obtenir un fichier XML de structure identique à celle du corpus d'apprentissage fourni par DEFT2010. Ce « corpus d'événements » se compose de 145 portions, une par année, chaque portion renfermant des expressions relatives à des événements ayant eu lieu dans cette année.

Voici, à titre d'exemple, le début de la portion correspondant à l'année 1837 :

```
<portion id="1837">
<meta>
<journal>WIKIPEDIA</journal>
<date annee="2010" mois="04" jour="02"/>
</meta>
<periode>1830</periode>
<texte>
6 juin : Assassinat du président Diego Portales au Chili par des militaires mutins.
6 septembre : Révolte de la Sabinada à Bahia, au Brésil (fin le 16 mars 1838).
2 octobre : Le Racer's storm, un des ouragans les plus puissants et les plus dévastateurs
...
Septembre : Révolte des Canadiens français, rapidement matée par les forces régulières britanniques.
6 novembre : Affrontement à Montréal entre l'Association patriote « Les Fils de la Liberté »
et les membres du « Doric Club » d'allégeance loyaliste. Saccage de maisons de patriotes.
16 novembre : Arrestation de chefs patriotes. Louis-Joseph Papineau réussit à se rendre aux États-Unis.
19 novembre : Manifestations de Patriotes à Québec.
...
</texte>
</portion>
```

3.2 Extraction des entités des corpus d'apprentissage

Les différentes sortes d'entités (entités nommées, termes complexes, noms communs, verbes) ont été extraites des corpus d'apprentissage par des outils propres à chaque type et cités ci-dessus. Lors de cette extraction nous avons conservé le lien entre chaque entité et toutes les classes (decennies, journaux) dans lesquelles elle a été rencontrée. Pour avoir une image de la distribution des entités dans les corpus d'apprentissage, nous avons calculé le nombre de *portion/articles* différents ($nbport$ pour la tâche n° 1 et $nbart$ pour la tâche n° 2) dans lequel apparaît chaque entité, ainsi que son nombre d'occurrences ($frequence$) dans la classe. Il arrive en effet qu'une entité apparaisse plusieurs fois dans un même *portion/article* ($nbport \leq frequence$ et $nbart \leq frequence$).

À partir de ces extractions nous avons construit des images des corpus d'apprentissage (le corpus de DEFT et elui construit à partir de *Wikipedia* pour la tâche n° 1 et le corpus de DEFT pour la tâche n° 2). Ces images prennent la forme de la liste des entités qu'ils renferment associées aux classes dans lesquelles elles apparaissent. Voici l'exemple de l'entité nommée *Croix Rouge* extraite du corpus *Wikipedia* où elle apparaît dans 7 portions réparties dans 6 décennies, et du terme *action terrestre* extrait du corpus d'apprentissage DEFT sur les origines où elle apparaît dans 3 articles appartenant à 2 journaux :

```
<entite type="EN" ressource="wikipedia">
   <lemme>Croix-Rouge</lemme>
   <decennies>
     <dec nbport='1' frequence='1'>1940</dec>
     <dec nbport='1' frequence='1'>1910</dec>
     <dec nbport='1' frequence='1'>1870</dec>
     <dec nbport='2' frequence='2'>1860</dec>
     <dec nbport='1' frequence='2'>1850</dec>
     <dec nbport='1' frequence='1'>1820</dec>
   </decennies>
 </entite>
```

Une entité nommée et ses décennies extraite du corpus *Wikipedia*

```
<entite type='TC' >
  <lemme>action terrestre</lemme>
  <journaux>
    <journal nbart='1' frequence='1'>D</journal>
    <journal nbart='2' frequence='3'>M</journal>
  </journaux>
</entite>
```

Un terme complexe et ses journaux extrait du corpus DEFT sur les origines

4 Recherche de marqueurs dans les corpus d'évaluation

Afin de déterminer de manière précise la classe la plus probable associée à une portion ou à un article, selon la tâche (section 5), il faut au préalable repérer, dans ces derniers, les entités apprises dans les corpus d'apprentissage (étape 1 de la figure 2).

4.1 Recherche des entités présentes dans le corpus d'évaluation

Pour chaque *portion/article* des corpus d'évaluation, nous avons repéré les différentes entités sur lesquelles nous avions décidé de travailler : les entités nommées (EN), les termes complexes (TC), les verbes (V) et les noms (N) avec les mêmes outils que ceux utilisés sur les corpus d'apprentissage. On obtient, pour chaque corpus d'évaluation, quatre listes (une par type d'entité) conservant la structure en *portion/article*, des entités qui en ont été extraites.

Voici par exemple quelques entités nommées d'une portion du corpus d'évaluation de la tâche n° 1 et quelques termes d'un article du corpus d'évaluation de la tâche n° 2 :

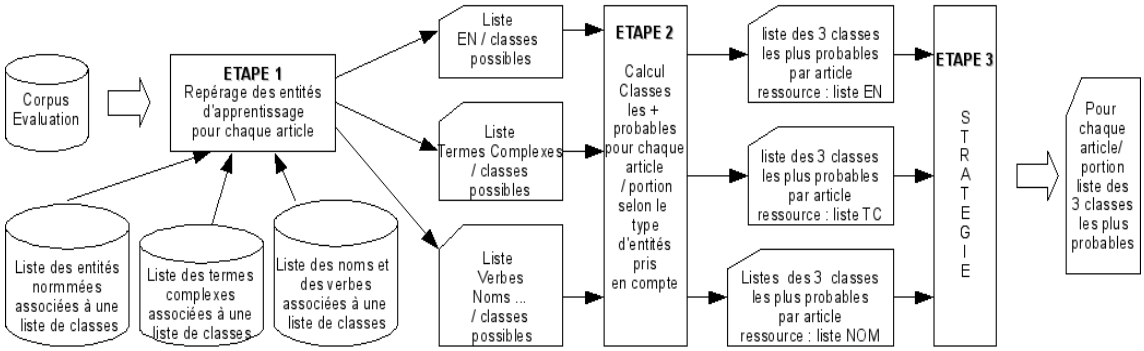


FIG. 2 – Détection de la classe d'une portion ou d'un article

```
<portion id="2143">
 ...
 <entite type="EN" nb="1">
  <lemme>Rodez</lemme>
 </entite>
 <entite type="EN" nb="1">
  <lemme>monts d' Aubrac</lemme>
 </entite>
 <entite type="EN" nb="1">
  <lemme>Louis-Philippe</lemme>
 </entite>
 ...
</portion>
```

Extrait du corpus d'évaluation de la tâche n° 1

```
<article id='1'>
   ...
    <entite type='TC' nb='1'>
      <lemme>jeu olympique de hiver</lemme>
    </entite>
    <entite type='TC' nb='1'>
      <lemme>médaille de or</lemme>
    </entite>
    <entite type='TC' nb='1'>
      <lemme>vache maigre</lemme>
    </entite>
    ...
</article>
```

Extrait du corpus d'évaluation de la tâche n° 2

4.2 Enrichissement des entités repérées dans les corpus d'évaluation par les données d'apprentissage

L'étape suivante consiste à enrichir les entités rencontrées dans les corpus d'évaluation par les différentes classes qui leur ont été associées lors de la phase d'apprentissage (étape 1 de la figure 2).

Voici les entités nommées enrichies de la portion de l'exemple précédent (tâche n° 1) :

```
<portion id="2143">
    ...
     <entite type="EN">
        <lemme>Rodez</lemme>
        <decennies>
           <dec nbport="1" ressource="appr">1890</dec>
           <dec nbport="1" ressource="appr">1850</dec>
           <dec nbport="1" ressource="appr">1910</dec>
           <dec nbport="1" ressource="appr">1930</dec>
           <dec nbport="1" ressource="wikipedia">1930</dec>
        </decennies>
     </entite>
     <entite type="EN">
        <lemme>monts d' Aubrac</lemme>
        <decennies/>
     </entite>
     <entite type="EN">
        <lemme>Louis-Philippe</lemme>
```

```
        <decennies>
           <dec nbport="1" ressource="appr">1870</dec>
           <dec nbport="2" ressource="appr">1830</dec>
           <dec nbport="1" ressource="appr">1840</dec>
           <dec nbport="1" ressource="wikipedia">1860</dec>
           <dec nbport="2" ressource="wikipedia">1840</dec>
           <dec nbport="2" ressource="wikipedia">1830</dec>
        </decennies>
     </entite>
    ...
  </portion>
```

Deux des trois entités nommées de la portion 2143 appartiennent aux listes d'entités nommées apprises avec le corpus d'apprentissage : "Rodez" et "Louis Philippe" : on enrichit ainsi ces deux entités par la liste des décennies probables en conservant pour chaque décennie, le nombre de portions (*nbport*)dans lesquels on a trouvé l'entité et l'information nous indiquant dans quel corpus d'apprentissage l'entité a été trouvée (*ressource*).

De même pour la tâche 2, voici les termes complexes de l'article de l'exemple précédent, enrichis des informations sur les origines ramenées pa l'apprentissage :

```
<article id="1">
   ...
   <entite type="TC">
       <lemme>jeu olympique de hiver</lemme>
       <journaux>
          <journal nbart="6">D</journal>
          <journal nbart="1">M</journal>
          <journal nbart="2">P</journal>
       </journaux>
    </entite>
    <entite type="TC">
       <lemme>médaille de or</lemme>
       <journaux>
          <journal nbart="15">D</journal>
          <journal nbart="6">E</journal>
          <journal nbart="4">M</journal>
          <journal nbart="15">P</journal>
       </journaux>
    </entite>
        <entite type="TC">
       <lemme>vache maigre</lemme>
       <journaux>
          <journal nbart="1">M</journal>
       </journaux>
    </entite>
 </article>
```

Ici tous les termes complexes repérés dans l'article 1 ont été enrichis car ils sont tous présents dans le corpus d'apprentissage. Pour la tâche n° 2, on garde, pour chaque journal concerné par une entité, le nombre d'articles (*nbart*) dans lequel a été repérée cette entité pour ce journal.

4.3 Analyse du corpus d'évaluation

Suite à l'enrichissement de certaines entités du corpus d'évaluation, nous avons voulu étudier de manière plus précise ce corpus :

– en évaluant le nombre d'entités enrichies présentes dans ce corpus pour voir la couverture de nos listes d'apprentissage

– en mesurant l'ambiguïté des classes associées à chaque entité

4.3.1 Tâche 1

Dans le corpus d'évaluation nous avons reconnu un certain nombre d'occurrences d'entités de chaque type (Nbre d'occ), dont un certain nombre enrichies par nos listes d'apprentissage (Nbre d'occ enrichies) :

Type	Nbre d'occ.	Nbre d'occ. enrichies	Rapport
Entités Nommées	17760	9756	54,93 %
Termes Complexes	37311	9232	24,74 %
Noms	105727	102305	96,76 %
Verbes	60158	59578	99,04 %
	220956	180871	81,86 %

Comme nous l'avons constaté dans la section 2.1, les types "entités nommées" et "termes complexes" semblent les plus pertinents pour déterminer la décennie d'une portion, puisqu'ils permettent de faire référence à des événements ayant lieu lors de la décennie. Toutefois, on constate qu'une entité nommée sur deux du corpus d'évaluation n'est pas présente dans les corpus d'apprentissage et qu'il en est de même pour un terme complexe sur quatre. La tâche 1 semble donc difficile à résoudre, par un manque de connaissances initiales. De plus les entités enrichies semblent loin d'être spécifiques à une décennie particulière au regard du pourcentage d'entités n'étant associées qu'à une seule décennies :

Nbre de décennies associées : nb	$nb = 1$	$2 \leq nb \leq 4$	$5 \leq nb \leq 10$	$nb > 10$
Entités Nommées	15,44 %	15,71 %	14,84 %	54,01 %
Termes Complexes	42,40 %	32,06 %	14,61 %	10,93 %

On constate que pour les entités nommées la tâche est d'autant plus complexe que 54,01 % des entités sont associées à plus de dix décennies. En regardant de plus près, on constate qu'il s'agit essentiellement des noms de régions, de pays, de villes.

4.3.2 Tâche 2

On réalise les mêmes calculs pour la tâche 2 :

Type	Nbre d'occ.	Nbre d'occ. enrichies	Rapport
Entités Nommées	38048	28089	73,83 %
Termes Complexes	89118	37762	42,37 %
Noms	180865	172066	95,14 %
Verbes	104576	103915	99,37 %
	412607	341832	82,85 %

Pour la tâche 2, le nombre d'occurrences d'entités nommées et de termes complexes augmente considérablement et devrait donc permettre de résoudre plus facilement la tâche. Mais on constate que parmi les occurrences de ces entités, beaucoup sont présentes dans plus d'un journal et pas forcément du même pays.

Journaux associés à une entité	d'un seul journal	seulement français	seulement québécois
Entités Nommées	17,64 %	14,09 %	20,37 %
Termes Complexes	36,98 %	24,25 %	26,37 %

Les deux tâches ne semblent pas simples à résoudre au vue des connaissances acquises avec le corpus d'apprentissage.

5 Résolution des tâches

Pour décider de la stratégie la plus adaptée pour chaque tâche, nous avons partitionné chaque corpus d'apprentissage en deux : un corpus test et un corpus d'apprentissage partiel.

Ainsi pour la tâche 1, nous avons constitué un corpus de test de 350 portions et un corpus d'apprentissage partiel de 2371 portions à partir duquel nous avons réalisé un apprentissage comme il a été décrit dans la section 3.

Pour la tâche 2, nous avons constitué de même un corpus de test de 370 articles et un corpus d'apprentissage partiel de 3349 articles où comme pour la tâche 1, ce dernier a servi dans la phase d'apprentissage pour ce test.

L'objectif était d'observer les résultats sur ce test et de déterminer quelles étaient les types d'entités à utiliser pour répondre au mieux à chacune des tâches : les entités nommées ? fusionnées à un ou plusieurs autres types ? ou de fusionner d'autres entités ?

Tous les résultats de cette section porteront sur les corpus test que nous avons fabriqués selon la description ci-dessus.

5.1 Calcul des classes les plus probables pour chaque portion/article

La deuxième étape, pour résoudre les différentes tâches consiste, à partir des entités enrichies, à déterminer les classes les plus probables pour chaque portion / article (voir étape 2 figure 2).

Ainsi pour chaque type d'entité, il s'agit de retourner la liste des classes susceptibles d'être la classe recherchée. Pour se faire, on fusionne toutes les classes de même type retournées par les entités enrichies.

Ainsi pour l'article 1 de la tâche 2 concernant les termes complexes, on obtiendra la liste suivante :

```
<article id="1">
   <journal nbentites="3" nbart="6">M</journal>
   <journal nbentites="2" nbart="21">D</journal>
   <journal nbentites="2" nbart="17">P</journal>
   <journal nbentites="1" nbart="6">E</journal>
 </article>
```

Cela signifie que dans l'article 1, ont été repérés 3 termes complexes issus de 6 articles du journal *Le Monde* du corpus d'apprentissage, 2 termes complexes dans 21 articles du journal *Le Devoir*, 2 termes complexes dans 17 articles du journal *La Presse* et 1 terme complexe dans 6 articles du journal *L'est républicain*.

Une fois la tâche de fusion pour chaque type d'entité réalisée, il faut classer ces différentes propositions. Plusieurs tests ont été effectués pour trouver le meilleur tri possible :

1. en fonction du nombre d'entités,
2. en fonction du nombre de portions (*nbport*) ou d'articles (*nbart*) du corpus d'apprentissage où ont été apprises les entités,
3. en fonction du nombre d'entités puis du nombre de portions (*nbport*) ou d'articles (*nbart*),
4. en fonction d'un rapport entre le nombre d'entités et le nombre de portions ou d'articles ...

Au final, le tri 3 est celui qui amène le plus grand nombre de bons résultats pour la tâche 1, donc pour chaque article dans chaque fichier correspondant à un type d'entité, les classes seront triées selon le nombre d'entités puis le nombre de portions ou d'articles.

Maintenant il s'agit de déterminer quel types d'entités utiliser pour résoudre les différentes tâches.

5.2 Stratégie

C'est au niveau de la stratégie que l'on peut noter une différence entre les deux tâches auxquelles nous avons participé.

5.2.1 Recherche d'une stratégie pour la tâche 1

Pour la tâche 1, la stratégie doit tenir compte du nombre important de décennies associées à chaque entité (comme nous l'avons vu dans la section 4.3), quel que soit le type d'entité.

Prenons notre exemple, pour la portion 2143, nous aurons pour le type d'entités EN le résultat suivant :

```
 <portion id="2143">
     <decennies>
        <dec nbentites="1" nbport="4" ressource="appr;wikipedia">1830</dec>
        <dec nbentites="1" nbport="2" ressource="appr;wikipedia">1930</dec>
        <dec nbentites="1" nbport="3" ressource="appr">1840</dec>
<dec nbentites="1" nbport="1" ressource="appr">1890</dec>
        <dec nbentites="1" nbport="1" ressource="appr">1850</dec>
        <dec nbentites="1" nbport="1" ressource="appr">1910</dec>
        <dec nbentites="1" nbport="1" ressource="appr">1870</dec>
        <dec nbentites="1" nbport="1" ressource="wikipedia">1860</dec>
      </decennies>
   </portion>
```

Pour la portion 2143, 8 décennies sont proposées et la décennie 1830 semble la décennie la plus probable pour la publication de la portion. Nous rappelons, en effet, que les décennies proposées sont triées par nombre d'entités puis par nombre de portions.

Plusieurs questions se posent ainsi à l'issue du calcul :
– Quels types d'entités utiliser pour avoir les meilleurs résultats ?
– Comment combiner les résultats obtenus pour chaque type d'entité si nous utilisons plusieurs entités pour résoudre la tâche ?
– Doit on prendre en compte, pour chaque type d'entités, toutes les décennies qui lui sont associées ou seulement les 3 meilleures ?

A partir du corpus test que nous avons extrait du corpus d'apprentissage, plusieurs tests ont été menés en faisant varier plusieurs paramètres (comme le tri des décennies, le nombre de décennies pris en compte lors de la combinaison, etc.). Nous présentons ci-dessous l'intervalle des précisions[2] calculées sur les différents tests, pour les différents types d'entités et leur combinaison.

Types Entités	Précision Minimale	Précision Maximale
EN	30,57 %	32,86 %
NOM	37,14 %	40 %
TC	29,14 %	32,86 %
VERBE	28,57 %	31,43 %
EN-TC-NOM-VERBE	36,57 %	42,86 %

D'autres combinaisons ont été testées mais c'est la combinaison des 4 types d'entités qui a donné les meilleurs résultats.

Pour les 3 runs que nous avons soumis, nous avons donc pris en compte tous les types d'entités (EN, TC, NOM et VERBE).

La différence entre les 3 runs porte d'une part sur la méthode de combinaison des résultats :
– RUN 1 : combinaison pour chaque portion des 3 premières décennies retournées par chaque type d'entité,
– RUN 2 et 3 : combinaison pour chaque portion de TOUTES les décennies retournées par chaque type d'entité.
d'autre part sur la manière de trier le résultat de la combinaison :

– RUN 1 et 2 : tri en fonction du nombre d'entités du corpus d'apprentissage présentes dans le corpus d'évaluation, puis du nombre de portions dans lesquelles ces entités étaient présentes dans le corpus d'apprentissage, pour cette décennie,
– RUN 3 : tri en fonction du nombre d'entités du corpus d'apprentissage présentes dans le corpus d'évaluation, puis du nombre d'entités ayant retourné la décennie en première position, puis du nombre de portions.

Avec le corpus test, la combinaison de TOUTES les décennies retournées par chaque type d'entité améliore les résultats, mais pas de manière flagrante d'où les deux runs proposés. Le taux de confiance d'une décennie pour une portion est égal au nombre de portions dans lesquelles ses entités ont été apprises, par rapport à la somme des portions dans lesquelles les entités des trois décennies retournées ont été apprises.

5.2.2 Recherche d'une stratégie pour la tâche 2

Pour déterminer le journal dans lequel est paru l'article, nous avons fait varier également les différents types d'entités pour définir les 3 journaux les plus probables par article. Nous avons ainsi calculé le pourcentage de bonnes réponses (voir le tableau ci dessous), retournées en première position (Pos1), en deuxième position (Pos2) et en troisième (Pos3).

[2]Nombre de portions retournant la bonne décennie dans les trois premières proposées / Nombre de portions

Types Entité	Pos1	Pos2	Pos3
EN	58,65 %	26,76 %	8,65 %
NOM	43,24 %	26,22 %	17,84 %
TC	55,95 %	27,03 %	11,89 %
VERBE	32,16 %	24,59 %	21,08 %
EN-TC	59,73 %	26,49 %	10,54 %
EN-NOM	55,68 %	28,11 %	9,46 %
TC-NOM	52,97 %	26,22 %	13,78 %
EN-TC-NOM	58,92 %	27,84 %	9,46 %
EN-TC-VERBE	59,73 %	25,68 %	10,54 %
EN-TC-NOM-VERBE	60,81 %	24,05 %	10,54 %

L'utilisation des entités nommées apprises par le biais du corpus d'apprentissage partiel semble indispensable au vu du pourcentage de bonnes réponses en première position (58,65 %), comme les termes complexes (55,95 %).

Ainsi pour nos 3 runs, nous choisissons les 3 stratégies suivantes :

1. Entités Nommées + Termes Complexes
2. Entités Nommées + Termes Complexes + Noms
3. Entités Nommées + Termes Complexes + Noms + Verbes

Les deux premières stratégies nous permettent en effet d'obtenir les meilleurs résultats quant au nombre de bonnes réponses parmi les 3 retournées, et la troisième stratégie correspond au meilleur taux de bonnes réponses en première position.

Il est évident qu'il aurait été aussi très intéressant d'évaluer la tâche 2 avec un seul type d'entité, notamment pour les entités nommées et les termes complexes, puisque les résultats sont similaires. Le tri des résultats des combinaisons de la tâche 2 est identique à celui des runs 1 et 2 de la tâche 1. Le taux de confiance en un journal pour un article est égal au nombre d'articles dans lesquels ses entités ont été apprises, par rapport à la somme des articles dans lesquels les entités des trois journaux retournés ont été apprises.

Comme nous l'avons dit au début de l'article, nous avons réalisé notre apprentissage sur le nom de journal. Pour déterminer le pays où a été publié l'article, nous nous basons sur les résultats obtenus pour la détermination du journal (3 propositions maximum). Le taux de confiance dans le pays est déterminé en fonction de la somme des taux de confiance des journaux qui lui sont associés (*Le Devoir* et *La Presse* pour le *Québec* et L'*Est Républicain* et *Le Monde* pour la France).

6 Évaluation des résultats et conclusion

6.1 Tâche 1

Nos résultats :

	Macro Rappel	Macro Précision	Macro F-mesure
Run 1	5,1 %	5 %	5 %
Run 2	5,3 %	5,2 %	5,3 %
Run 3	5,3 %	5,2 %	5,3 %

À la lecture de quelques portions du corpus d'apprentissage, nous n'avons pas relevé de marques linguistiques spécifiques à une période particulière, encore moins à une décennie. C'est pourquoi la seule méthode qui nous a semblé possible était de repérer des éléments de lexique dans le corpus d'apprentissage.

Au vu de la moyenne des résultats obtenus par l'ensemble des participants, le corpus d'apprentissage semble ne pas renfermer suffisamment d'informations pour dater les portions du corpus d'évaluation. Il aurait d'autre part été intéressant de pouvoir travailler sur des données non dégradées par l'OCR car nous n'avons pas eu le temps de mesurer l'impact de ces erreurs.

En ce qui concerne notre travail nous sommes conscientes qu'il nous aurait fallu mieux cibler chaque catégorie d'entité :
– ne conserver comme entités nommées que celles qui sont des vrais marqueurs d'événements c'est à dire les noms propres de personnes, de lieux (typiques comme, par exemple, *jardin public* mais non géographiques), d'événements (comme par exemple *Exposition Universelle* ou *Jeux Olympiques*)
– conduire un travail de réflexion approfondi sur la nature des termes complexes à conserver,
– ne conserver que certaines catégories de noms et de verbes, ce qui aurait nécessité d'autres ressources externes.

6.2 Tâche 2

	Rappel	Précision	F-mesure
* Run 1 - Pays	72,1 %	72,5 %	72,3 %
Run 1 - Journal	41,9 %	43 %	42,5 %
Run 2 - Pays	69,2 %	69,5 %	69,4 %
Run 2 - Journal	39,6 %	41,3 %	40,4 %
Run 3 - Pays	68,5 %	68,8 %	68,7 %
Run 3 - Journal	39,3 %	41,4 %	40,3 %

De manière analogue nous n'avons pas trouvé dans le corpus d'apprentissage de marques linguistiques spécifiques à la langue française du Québec ou de la France. Ceci nous a donc conduites vers des méthodes analogues pour les deux tâches.

Toutefois le corpus d'apprentissage a constitué une ressource mieux adaptée à la tâche demandée.

Comme nous l'avons constaté sur notre corpus test, ce sont les entités nommées et les termes complexes qui permettent d'obtenir les meilleurs résultats. Comme pour la tâche précédente il aurait fallu mieux cibler chaque catégorie d'entités.

Nous avons ramené la tâche à la reconnaissance des journaux, à cause de la dépendance pays, journal. Il serait sans doute intéressant de tester une stratégie en deux temps :
– apprentissage puis reconnaissance du pays indépendamment du journal,
– apprentissage puis reconnaissance du journal sur des corpus réduits aux pays.
Nous avons été très intéressées par ce travail d'investigation. Nous pensons qu'il y a encore beaucoup de travail a effectuer sur chacun de ces thèmes, tant dans la construction des ressources d'apprentissage que dans la mise en œuvre de nos méthodes.

Références

DAILLE B. (2003). Information Extraction in the Web Era. In M. PAZIENZA, Ed., *Terminology Mining*, p. 29–44. Springer.

FOUROUR N. (2002). Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. In J.-M. PIERREL, Ed., *Actes de TALN 2002 (Traitement automatique des langues naturelles)*, p. 265–274, Nancy : ATALA ATILF.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing(NeMLaP-1)*, p. 44–49.

L'Approche MARF à DEFT 2010: A MARF Approach to DEFT 2010

Serguei A. Mokhov
Université Concordia University
Montréal, QC, Canada
mokhov@cse.concordia.ca

Résumé. On présent l'approche MARF aux problèmes de classification de variation diachronique et origine géographique des textes de la presse francophone pour l'atelier DEFT 2010. Cet étude utilise MARF, un framework open-source écrit en Java pour la reconnaissance des formes automatique en générale, incluant le traitement multimédia, l'analyse forensique de fichiers, la reconnaissance des auteurs, des langues parlées (accents), des sexes, de l'âge, et langues naturelles. On présent les officiels et les meilleures résultats obtenus et notre approche, ses difficultés, avantages et désavantages etc. Pour les résultats complets veuillez consulter un autre document relié (Mokhov, 2010a).

Abstract. We present a MARF-based approach to classification problems of the decades and place of origin of various French publications in the DEFT 2010 challenge. This case study of MARF, the open-source Java-based Modular Audio Recognition Framework, is intended to show the complete general pattern recognition pipeline design methodology for machine learning to test and compare multiple algorithms, including supervised and unsupervised, statistical, etc. learning and classification for a spectrum of recognition tasks, applicable not only to audio recognition but to general pattern recognition for various NLP applications, e.g. writer, language identification, and others. We summarize our best results and the results used for the challenge here along with the methodology used to obtain them. For vast, a lot more complete results of this work please refer to the related document (Mokhov, 2010a).

Mots-clés : DEFT2010, MARF, frameworks, comparaison des algorithmes pour TAL.

Keywords: DEFT2010, MARF, frameworks, algorithm comparison for NLP.

1 Introduction

We present an approach to the DÉfi Fouille de Textes (DEFT) 2010 challenge (Forest *et al.*, 2010) by using the MARF framework (The MARF Research and Development Group, 2002 2010; Mokhov *et al.*, 2002 2003; Mokhov, 2008a,b; Mokhov & Debbabi, 2008; Mokhov *et al.*, 2009; Mokhov, 2010b; Mokhov & Vassev, 2009) by combining approaches from the related MARF applications into `DEFT2010App`. The DEFT2010 NLP challenge proposed two tracks in identification within francophone press : Piste 1 for identification of the decade of a publication, and Piste 2 for publications varying across geographic locations, specifically France vs. Quebec from several prominent journals. The corpora were compiled by Cyril Grouin from a variety of sources kindly provided by Gallica, CEDROM-SNi, ELDA, and CNRTL (Grouin, 2010a,c,e,d,b).

To the author's knowledge MARF is a still holding up as a unique framework that attempts various signal processing, etc. and NLP for comparative studies of implementations of algorithms and algorithm combinations. The closest open-source system is probably CMU Sphinx (The Sphinx Group at Carnegie Mellon, 2007 2010), which is a powerful speech-to-text system, but at the same time too complex for comparative scientific experiments MARF's primarily designed for. Plus, MARF's multifaceted approach allowed it to be used outside of the domain of audio and voice processing (Mokhov, 2010c).

The core founding works for this approach are (Mokhov, 2008a,b; Mokhov *et al.*, 2009; Mokhov, 2010c) adapted to the NLP tasks using the same multi-algorithmic approach – providing a framework for algorithm selection and selecting the best available combination of algorithms for a given task (Mokhov, 2010c).

Some MARF's example applications (on which `DEFT2010App` is based), such as text-independent speaker-identification, language (natural and programming) identification, natural language probabilistic parsing, etc. are released along with MARF as open-source and some are discussed in several related publications mentioned earlier (Mokhov *et al.*, 2002 2010; Mokhov & the MARF Research & Development Group, 2003 2010a,2; Mokhov, 2008 2010).

2 Methodology

Using MARF, we show the usefulness of helping researchers to decide the algorithm combinations the best or better suited for each particular task they need to work on (Mokhov, 2008b, 2010c). MARF provides an ability to train the system for each task and then test it on the unseen samples giving back statistics for each algorithm combination (different permutations of algorithms are used in loading, preprocessing, feature extraction, and classification stages) used from the best to the worst, including second-best statistics augmented with statistical estimators and NLP parsing and other modules. The DEFT workshop participants (and later the rest of the world) will get the complete set of non-copyrighted materials and will be able to extend it and contribute to the project, which is open-source, if they choose to. This approach is similar to that described in (Mokhov, 2008b), but applied to corpora instead of voice samples using primarily spectral analysis of texts (spectral analysis was previous applied to texts by others as well, e.g. (Vaillant *et al.*, 2006), and speech and signal processing (Russell & Norvig, 1995; Press, 1993; O'Shaughnessy, 2000; Ifeachor & Jervis, 2002; Bernsee, 1999 2005)).

2.1 Two Pipelines Approach / L'Approche Deux Pipelines

This approach is applied to both tracks, Piste 1 for the decades and Piste 2 for the geographic locations. We detail the methodology further.

2.1.1 Classical MARF Pipeline Approach

The classical MARF pipeline is in Figure 1 (Mokhov, 2008b). It shows variety of algorithms that can be selected at run time via configuration settings to allow any permutation of the algorithms on the recognition path. The pipeline is augmented with the classical statistical NLP components from (Jurafsky & Martin, 2000; Martin, 2003) and others to do the statistical NLP processing are illustrated in a high-level UML class diagram in Figure 2. It's split into the subframeworks covering n-gram language models, natural language parsing, collocations, etc. and the supporting utility modules. The focus here, however, is primarily on the statistical analysis and recognition using machine learning, language models, and signal processing techniques.

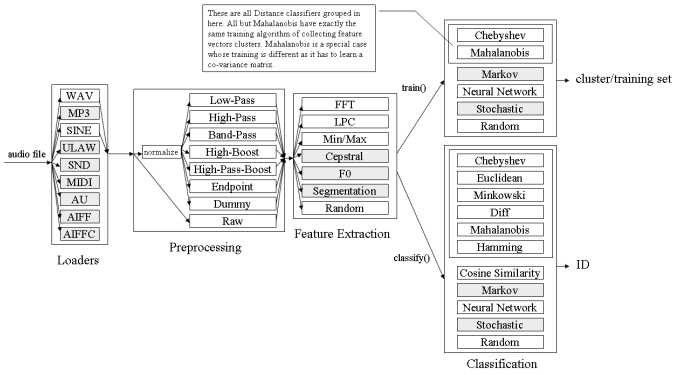


FIGURE 1 – Classical Pattern Recognition Pipeline of MARF

2.1.2 NLP MARF Pipeline Approach

This is another, somewhat distinct branch of experiments than that briefly described in Section 2.1.1 (while still using the same MARF-implementing system). This path of experiments is based on the largest part of the `LangIdentApp` (Mokhov & the MARF Research & Development Group, 2003 2010a) and the corresponding `MARF.NLP` class as well as the statistical estimators framework shown on Figure 2. This approach takes a different set of options and the parameters than the one in Section 2.1.1. We tokenize the input stream as individual characters and the build classical n-gram models with $n = 1, 2, 3$ and the corresponding statistical and smoothing estimators. The language model is then the smoothed using a 1D, 2D, or 3D frequency matrix and any future comparison for classification is that of the matrices learned. Each portion of an article (for both tasks) is used from the training data to compute the language models and serialize them per estimator. Then, we use the same models to test on the testing data. The precision statistics is computed identically to the classical pipeline approach. Thus, we still have a comparison of algorithms in a pipeline, but this is a more traditional statistical NLP pipeline.

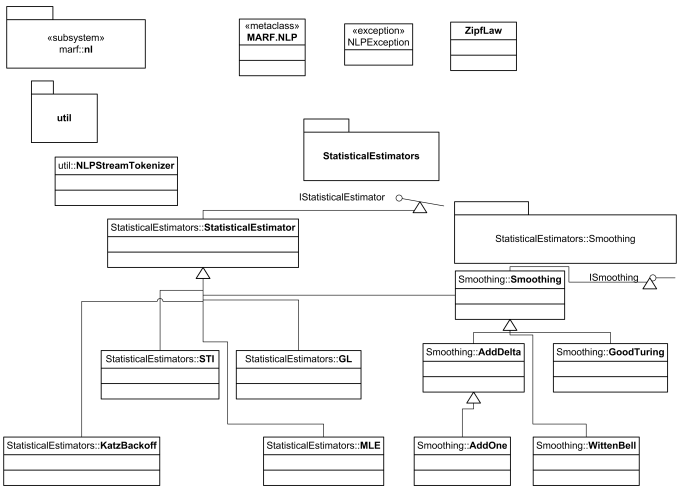


FIGURE 2 – A Partial Set of the NLP Components of MARF

2.2 Variable Tunable Parameters

This section describes the permutations of variable parameters used throughout the experiments to determine the best permutation/combination available.

2.2.1 DEFT'10-Specific Options

These options combine with all the subsequent options to select an algorithm combination to do the actual data processing. What follows is to tell what type of training and testing data to process in the pipelines.

1. `-piste1` – indicates we are dealing with the Piste 1 data, such that the internal data structures are properly configured to handle that ; it corresponds to Section 2.6.
2. `-journal` – a Piste 2 option indicating to use journals as primary classes as opposed the geographical locations ; and then compute the latter from the former (as knowing each journals gives uniquely the place of origin, while the reverse is not true). It corresponds to Section 2.7.3.
3. `-text-only` indicates to process just the article fragments themselves. This is default of loading bodies of the articles.
4. `-title-only` – a Piste 2 option indicating to use only article titles for the tasks instead of the article bodies.
5. `-title-text` – a Piste 2 option indicates to use both the article title and its body as a sample It corresponds to Section 2.7.2.
6. `-ref` – indicates to load and validate against the reference data supplied by the organizers (added later when the data became available)

2.2.2 Sample Loading and Interpretation

At present there are two types of interpretation of the data performed after the initial XML loading : (a) interpreting each character bigram as an amplitude of a waveform and treat the whole input data as an audio wave signal with 2 bytes per amplitude value, mono, WAVE PCI-encoded data (all are the defaults in MARF). The `Sample` objects produced in that way are processed by the traditional MARF pipeline, which is more signal processing and spectral oriented (cf. Figure 2.1.1). (b) interpreting each n-gram ($n = 1, 2, 3$) token from `NLPStreamTokenizer` to compute the language models and use the NLP pipeline of MARF (cf. Section 2.1.2). `NLPStreamTokenizer` is itself configurable to filter different kinds of tokens and characters. Reduction of the data set loaded for training is another variable to explore as sometimes "oversaturation" of the training models with a lot of training data actually lower the recognition performance. There are a lot more experiments possible with the loaders if the interpreted data, which was not done for this submission, but may be performed later in one of the versions of (Mokhov, 2010a).

2.2.3 Preprocessing

Preprocessing algorithms in the classical pipeline take loaded `Sample` objects of arbitrary length and produce arbitrary length "preprocessed" objects of the same type. The resulting objects may become smaller or keep the same length as the original depending on the module used. These are typically normalization and filtering modules. They also can be chained, but the chaining aspect was not really fully explored in this work.

1. `-silence` and `-noise` – suppress "silence" and "noise" from the data (cf. Section 2.7.4).
2. `-norm` – whether to apply normalization of the input data or with `-raw` just to pass it through without any preprocessing.
3. `-low`, `-high`, `-band`, `-bandstop`, `-high`, `-boost`, `-highpassboost` – are a variety of FFT-based filters, such as low-pass, high-pass, band-pass, band-stop, high-frequency amplitude boost, and high-pass with high frequency amplitude boost.

2.2.4 Feature Extraction

All feature extractor modules take a variable-length sample input from preprocessing and produce a fixed-length feature vector x. The resulting feature vectors get stored in various training models such as mean (default) or median clusters, or plain collections of feature vectors.

1. `-fft` – use the Fast Fourier Transform (FFT) with 512 default features (frequencies)
2. `-lpc` – use the Linear Predictive Coding (LPC) algorithm with the default 20 features (poles)
3. `-aggr` – uses aggregation of the FFT and LPC features as a feature vector x
4. `-minmax` – use 50 minimums and 50 maximums by default as features

2.2.5 Classification

Classical mostly spectral pipeline. Classifier modules typically take the fixed-length feature vectors x from the feature extractors and then store them as either mean (default) or median cluster or just as a collection of feature vectors came in during training. This is true for the majority of distance and similarity classifiers. Additional training is done for Mahalanobis distance (Mahalanobis, 1936) (covariance matrix) and the artificial neural network (the network itself). These respresent the language models that are compared against when testing.

1. `-cheb`, the Chebyshev (aka city-block or Manhattan) distance classifier (Abdi, 2007)
2. `-eucl`, the Euclidean distance
3. `-cos` – use cosine similarity measure as a classifier (Garcia, 2006; Khalifé, 2004)
4. `-mink` – use Minkowski distance measure
5. `-diff` – use Diff distance measure (Mokhov, 2008b)
6. `-hamming` – use Hamming distance measure (Hamming, 1950)

Mean cluster vs. median cluster vs. feature set. It has been shown that the choice of cluster or absence thereof may positively impact precision (Mokhov, 2008b) and providing distinct top algorithm combinations from that of mean. Both type of clustering while save processing time and storage space, are deemed to contribute less accurately to precision than just keeping all the collections of all feature vectors as-is. Thus, this set of experiments is to test variability of the precision (and other metrics) to the Piste 1 and Piste 2 tasks. These experiments can be combined with the previously described variations and the mean clustering was the default that was used. The follow up experiments and observations will be reported in the subsequent ongoing study in (Mokhov, 2010a).

NLP statistical estimators pipeline. In this pipeline we script the task in a similar manner where `-char` used in the preprocessing to set the character model for the tokenizer; `-unigram`, `-bigram`, and `-trigram` act as feature selectors, and finally the smoothing estimators enumerated below act as classifiers. This experiment is still ongoing and its results are expected to appear in (Mokhov, 2010a).

1. `-add-delta` by implementing the general Add-Delta smoothing we get MLE, ELE, and Add-One "for free" as special cases of Add-Delta :
   – $\delta = 0$ is MLE (maximum likelihood estimate), `-mle`
   – $\delta = 1$ is Add One, `-add-one`
   – $\delta = 0.5$ is ELE, `-add-delta`
2. Witten-Bell and Good Turing smoothing options `-witten-bell` and `-good-turing`. These two estimators were implemented as given in hopes to get a better recognition rate over the Add-Delta family.

2.3 Omitted "Slow" and Other Algorithms of MARF

The following algorithms were not conclusively tested yet for either task and are either in process of execution or being debugged due to slowness, etc. Their respective results are to appear in (Mokhov, 2010a) as they become available. Please consult that article from time to time for the respective updates. The algorithms include the artificial neural network, continuous fraction expansion (CFE) filters (Haridas, 2006) (high-, low-, band-pass, and band-stop), Zipf's Law, and Mahalanobis distance – all combined with all the mentioned algorithms as applicable.

2.4 Zipf's Law

The `ZipfLaw` classifier exists for both classical and NLP pipelines of MARF. In the former it uses the `DiffDistance` class to compute the distance between the two ranked dictionaries of `Double` values. This corresponds to the `-zipf` option in the `DEFT2010App` application. This is a very slow approach by default in the traditional pipeline. In the NLP pipeline, the `ZipfLaw` module is used to collect the most discriminative terms from training and rank them. We keep the top N rank lexemes (n-grams, where lexemes can be a combination of n characters or words producing shorter or longer dictionaries). On classification, we compute a ranked ZipfLaw top N rank set, and compute their distance or similarity as vectors to decide the class. The option `-cheat` corresponds to this. (At the time of this writing we do not have the Zipf's Law results available ; all options are enumerated in (Mokhov, 2010a) and that's where we plan placing the Zipf's law results as well).

2.5 Extra Testing on the Unseen Data

This section proposes an additional testing methodology on the unseen data while the reference and the test data are not available.

2.5.1 Using Piste 1's Data in Piste 2

All of the Piste 1's data are reusable in Piste 2 to partially test Piste 2's setup on an unseen corpora, which is conveniently and freely available in order to improve the testing of the system. This extra testing requires to be able to load both the training data of the Piste 1 and the language models of Piste 2 at the same time and provide the appropriate option and scripting support as well as mapping the relevant Piste 1's journals

to France. While only partial testing, it is more honest than testing on the training data. This methodology is further being developed in (Mokhov, 2010a).

2.5.2 Using Francophone Websites

Using the `wget` (Niksic & Free Software Foundation, Inc., 1996 2009) tool in a scripted manner on the French and Quebec web sites in a recursive manner helps creating a corpora of HTML pages. Then, treat either each page as an article of its own (after tags removal) or each textual node paragraph larger than 2000 bytes as an article. The good candidate sites for this type of testing data are various government and press sites openly available. The results of this testing are planned to further to appear in (Mokhov, 2010a).

2.6 Piste 1 : Decades

Decades proved difficult to get correctly. The best configuration on the training and testing data are in Section 3.2. Most of the general methodology described in Section 2.1 and Section 2.2 were used to come up with the best available configuration to date that produces highest macro precision. Top N (usually $N = 50$) of those best configurations were subsequently measured with the testing and reference data later provided by the organizers. The complete result set for Piste 1 trials can be found in (Mokhov, 2010a).

2.7 Piste 2 : Place of Origin

There is a lot more variety in the experiments for the methodology used in Piste 2 as opposed to Piste 1. The major variations are listed here. The results are summarized in Section 3.3.

2.7.1 Small Increase in the Article Text Size

We collect titles as well as texts. For the training purposes one way to experiment is to merge the two effectively increasing a little bit the training sample size for each article fragment. The experiment is to verify if adding titles to texts increases the precision of detection. The results prove that this hypothesis is not quite correct. In fact, the precision has gone worse and the processing time increased in many configurations.

This experiment was conducted both with the location being the leading class, followed by the journal as well as when the journal is the leading class as described in Section 2.7.3.

2.7.2 Titles vs. Texts

For a 300-word article, its title may act as an abstract. Often in NLP, e.g. BioNLP, NLP techniques area applied to the article abstracts rather than to the complete texts. A title in our sample can be considered as an abstract on the same scale of a short article excerpt. Thus, we test the approach to see if titles alone are sufficient to generate enough precision and increasing the performance.

Disadvantages : sometimes articles do not have titles (i.e. `<titre />`, so only guessing at random (an improvement is to use the main body in such cases).

– Training data's articles without titles : 297
– Testing data's articles without titles : 207

Despite the missing titles in some articles and without the implementation of the mentioned improvement, the results produces the higher precision than in the other experiments mentioned thus far.

2.7.3 Journal as a Primary Class

There is a relationship between journals and locations where they were published. The journals uniquely identify the place ; however, the place does not necessarily uniquely identify the journal, but puts a constraint on what the journals may be allowing cross-validation to prevent journal-place mismatches.

In most of the experiments the location was a leading class, with the journal being identified later as follows :

– Assume a common journal per location without actually running a classifier.

– Run the pipeline the second time for the journal class once the location is known and pick the likely journal from that country.
– Assume the journals are the primary class and then deduce the location from them. It turns out while the journal is primary, despite relatively low macro precision for journals ($\approx 40\% - 48\%$), the macro precision for some experiments turned out better for countries and for an equivalent non-journal identification.

2.7.4 Silence and Noise Removal

These are spectral options (`-silence`, `-noise`) from the classical MARF. The experiments show the default options of signal or noise removal (see Section 2.2) made difference some cases. The silence is removed when the preprocessed wave signal's amplitude points are below a certain empirical threshold (0.001). The "noise" is presently removed by applying the low-pass FFT filter by default. The complete result set containing the experiments along with others are in (Mokhov, 2010a).

3 Results / Résultats

Here we present some official results first, and then other related experiments and improvement upon them. These are just a small fraction of all the experiments conducted to date.

3.1 Official / Officiels

Our official results at the submission time were not great ; a part of the reason the testing on all combinations were not completed yet (some are still executing), so the best available configurations at that moment were used, which are *emphasized* in the tables. Those configurations gave highest macro precision on the training data back then, found in Table 1 and Table 5. Those submissions produced the results found in Table 9 and and Table 10. We, however, did not stop the experiments running and conducting some more at the time of this writing, that are being maintained more-or-less in full in (Mokhov, 2010a). We were able to improve the results over the official submission, and what follows are some of the top picks.

TABLE 1 – Piste 1 : Top 10 configurations tested on the training data.

Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	*-piste1 -norm -fft -cos*	1025	2569	*28.52*
2	1st	-piste1 -silence -norm -aggr -cos	1025	2569	28.52
3	1st	-piste1 -silence -norm -fft -cos	1025	2569	28.52
4	1st	-piste1 -norm -aggr -cos	1025	2569	28.52
5	1st	-piste1 -raw -fft -cos	1023	2571	28.46
6	1st	-piste1 -silence -raw -aggr -cos	1023	2571	28.46
7	1st	-piste1 -silence -noise -raw -fft -cos	1023	2571	28.46
8	1st	-piste1 -noise -raw -fft -cos	1023	2571	28.46
9	1st	-piste1 -silence -noise -raw -aggr -cos	1023	2571	28.46
10	1st	-piste1 -noise -raw -aggr -cos	1023	2571	28.46

3.2 Piste 1 : Decades

The resulting tables pertinent to the Piste 1 are in Table 1 and Table 2 per configuration and per class with macro precision tested on the training data ; Table 3 and Table 4 correspond to the same with the testing and reference data, while in Table 9 is the official result at the submission time.

3.3 Piste 2 : France vs. Quebec

The resulting tables pertinent to the Piste 2 are in Table 10, Table 5, Table 6, Table 7, Table 8. Additional extra results that were obtained after the submissions are in Table 11, which shows consolidated top 4 macro precision results per configuration for the title + text experiment (cf. Section 2.7.1), title-only (cf. Section 2.7.2) ; then repeat of the three experiments with journal being the leading class (cf. Section 2.7.3).

TABLE 2 – Piste 1 : Macro precision on training data per decade across 834 configurations

Run #	Guess	Decade	GOOD	BAD	Precision,%
1	1st	1830	84722	125446	40.31
2	1st	1940	38330	147652	20.61
3	1st	1810	38811	171357	18.47
4	1st	1820	29107	181061	13.85
5	1st	1900	15229	166583	8.38
6	1st	1920	15360	168954	8.33
7	1st	1850	16807	193361	8.00
8	1st	1880	15643	194525	7.44
9	1st	1800	14974	195194	7.12
10	1st	1860	14038	195296	6.71
11	1st	1870	14037	196131	6.68
12	1st	1930	11314	173000	6.14
13	1st	1840	12468	197700	5.93
14	1st	1890	10727	176089	5.74
15	1st	1910	10408	173072	5.67
16	2nd	1830	105137	105031	50.03
17	2nd	1940	53906	132076	28.98
18	2nd	1810	98904	111264	47.06
19	2nd	1820	45219	164949	21.52
20	2nd	1900	28690	153122	15.78
21	2nd	1920	30199	154115	16.38
22	2nd	1850	32345	177823	15.39
23	2nd	1880	36923	173245	17.57
24	2nd	1800	38997	171171	18.56
25	2nd	1860	29063	180271	13.88
26	2nd	1870	28356	181812	13.49
27	2nd	1930	23456	160858	12.73
28	2nd	1840	25489	184679	12.13
29	2nd	1890	21281	165535	11.39
30	2nd	1910	19657	163823	10.71

TABLE 3 – Piste 1 : Testing on the evaluation + reference data top 10 configurations

Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	**-piste1 -ref -silence -bandstop -aggr -cos**	331	2390	**12.16**
2	1st	-piste1 -ref -noise -raw -aggr -eucl	315	2406	11.58
3	1st	-piste1 -ref -silence -raw -aggr -eucl	315	2406	11.58
4	1st	*-piste1 -ref -norm -fft -cos*	315	2406	*11.58*
5	1st	-piste1 -ref -raw -aggr -eucl	315	2406	11.58
6	1st	-piste1 -ref -noise -raw -fft -cos	315	2406	11.58
7	1st	-piste1 -ref -silence -raw -fft -eucl	315	2406	11.58
8	1st	-piste1 -ref -silence -raw -aggr -cos	315	2406	11.58
9	1st	-piste1 -ref -silence -noise -raw -fft -eucl	315	2406	11.58
10	1st	-piste1 -ref -silence -noise -raw -aggr -cos	315	2406	11.58

Table 12 lists the corresponding macro results per class. Following that, the tables Table 13, Table 14, Table 15, Table 16, and Table 17 list the corresponding output from the `3_evalueResultats_t2.pl` evaluation tool provided by the DEFT2010 organizers in the end each corresponding to the top configuration of each experiment.

3.4 Results Summary / Résumé des résultats

– Total more than 8688 configurations tested.
– Apparent highest precision and recall results come from title-only processing for Piste 2 despite some testing and training items missing titles.
– While journal-leading-class experiments give 48% at their best macro precision per configuration they sometimes give better results for the corresponding location, than if the location is the leading class.

TABLE 4 – Piste 1 : Macro precision on testing + reference data per decade across 49 configurations

Run #	Guess	Decade	GOOD	BAD	Precision,%
1	1st	1940	3852	5850	39.70
2	1st	1830	2187	6094	26.41
3	1st	1820	1571	6710	18.97
4	1st	1810	1180	7101	14.25
5	1st	1880	757	7524	9.14
6	1st	1900	864	9083	8.69
7	1st	1920	703	9097	7.17
8	1st	1850	554	7727	6.69
9	1st	1870	417	7864	5.04
10	1st	1840	413	7868	4.99
11	1st	1860	395	7935	4.74
12	1st	1930	442	9358	4.51
13	1st	1890	403	9250	4.17
14	1st	1800	281	8000	3.39
15	1st	1910	290	9559	2.94
16	2nd	1940	4828	4874	49.76
17	2nd	1830	3457	4824	41.75
18	2nd	1820	3068	5213	37.05
19	2nd	1810	1772	6509	21.40
20	2nd	1880	1770	6511	21.37
21	2nd	1900	1924	8023	19.34
22	2nd	1920	1725	8075	17.60
23	2nd	1850	1068	7213	12.90
24	2nd	1870	881	7400	10.64
25	2nd	1840	966	7315	11.67
26	2nd	1860	884	7446	10.61
27	2nd	1930	1150	8650	11.73
28	2nd	1890	806	8847	8.35
29	2nd	1800	978	7303	11.81
30	2nd	1910	554	9295	5.62

TABLE 5 – Piste 2 : Top 38 of 839 results of testing on the training data using text only

Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	**-silence -high -fft -cos**	2203	1516	**59.24**
2	1st	-silence -high -aggr -cos	2202	1517	59.21
3	1st	-high -fft -cos	2169	1550	58.32
4	1st	-high -aggr -cos	2167	1552	58.27
5	1st	-silence -band -aggr -cos	2154	1565	57.92
6	1st	-silence -band -fft -cos	2154	1565	57.92
7	1st	-silence -noise -band -fft -cos	2138	1581	57.49
8	1st	-silence -noise -band -aggr -cos	2134	1585	57.38
9	1st	-silence -high -aggr -eucl	2134	1585	57.38
10	1st	-silence -high -fft -eucl	2133	1586	57.35
11	1st	-high -fft -diff	2129	1590	57.25
12	1st	-high -aggr -diff	2127	1592	57.19
13	1st	-high -aggr -cheb	2126	1593	57.17
14	1st	-high -fft -cheb	2124	1595	57.11
15	1st	-band -aggr -eucl	2124	1595	57.11
16	1st	-band -fft -eucl	2123	1596	57.09
17	1st	-noise -band -aggr -cos	2122	1597	57.06
18	1st	-band -fft -cos	2122	1597	57.06
19	1st	-band -aggr -cos	2122	1597	57.06
20	1st	-noise -band -fft -cos	2121	1598	57.03
21	1st	-high -aggr -eucl	2119	1600	56.98
22	1st	-high -fft -eucl	2118	1601	56.95
23	1st	-silence -band -fft -eucl	2115	1604	56.87
24	1st	-silence -band -aggr -eucl	2113	1606	56.82
25	1st	-silence -band -aggr -diff	2084	1635	56.04
26	1st	-silence -band -aggr -cheb	2084	1635	56.04
27	1st	-silence -band -fft -cheb	2083	1636	56.01
28	1st	-silence -band -fft -diff	2082	1637	55.98
29	1st	-band -fft -cheb	2071	1648	55.69
30	1st	-band -aggr -cheb	2070	1649	55.66
31	1st	-noise -raw -lpc -eucl	2058	1661	55.34
32	1st	-silence -raw -lpc -eucl	2058	1661	55.34
33	1st	-silence -noise -raw -lpc -eucl	2058	1661	55.34
34	1st	-band -fft -diff	2058	1661	55.34
35	1st	-silence -norm -lpc -eucl	2058	1661	55.34
36	1st	-norm -lpc -eucl	2058	1661	55.34
37	1st	-band -aggr -diff	2058	1661	55.34
38	1st	*-raw -lpc -eucl*	2058	1661	*55.34*

TABLE 6 – Piste 2 : Macro precision across 839 results per location on training data, text only

Run #	Guess	Location	GOOD	BAD	Precision,%
1	1st	Quebec	932695	737754	55.83
2	1st	France	688020	761772	47.46

TABLE 7 – Piste 2 : Top 15 configurations on the evaluation and reference data, text only

Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	-text-only -ref -noise -band -fft -diff	1392	1090	56.08
2	1st	-text-only -ref -noise -band -aggr -diff	1391	1091	56.04
3	1st	-text-only -ref -noise -band -fft -cheb	1386	1096	55.84
4	1st	-text-only -ref -noise -band -aggr -cheb	1386	1096	55.84
5	1st	-text-only -ref -high -fft -diff	1380	1102	55.60
6	1st	-text-only -ref -band -aggr -cheb	1380	1102	55.60
7	1st	-text-only -ref -band -fft -cheb	1379	1103	55.56
8	1st	-text-only -ref -high -fft -cheb	1378	1104	55.52
9	1st	-text-only -ref -high -aggr -diff	1378	1104	55.52
10	1st	-text-only -ref -high -aggr -cheb	1377	1105	55.48
11	1st	-text-only -ref -band -aggr -diff	1376	1106	55.44
12	1st	-text-only -ref -band -fft -diff	1375	1107	55.40
13	1st	-text-only -ref -silence -band -aggr -cheb	1325	1157	53.38
14	1st	-text-only -ref -silence -band -fft -cheb	1322	1160	53.26
15	1st	-text-only -ref -silence -band -aggr -diff	1317	1165	53.06

TABLE 8 – Piste 2 : Macro precision across all configurations per location on evaluation data, text only

Run #	Guess	Location	GOOD	BAD	Precision,%
1	1st	France	32445	24052	57.43
2	1st	Quebec	31720	33401	48.71

TABLE 9 – Fichier évalué : `equipe_3_tache_1_execution_1_1006060550.xml`

```
- classe 1800 (attendus = 169, ramenes = 127.00, corrects = 9.00)
  rappel = 0.053 precision = 0.071 f-mesure = 0.061
- classe 1810 (attendus = 169, ramenes = 249.00, corrects = 32.00)
  rappel = 0.189 precision = 0.129 f-mesure = 0.153
- classe 1820 (attendus = 169, ramenes = 258.00, corrects = 29.00)
  rappel = 0.172 precision = 0.112 f-mesure = 0.136
- classe 1830 (attendus = 169, ramenes = 354.00, corrects = 36.00)
  rappel = 0.213 precision = 0.102 f-mesure = 0.138
- classe 1840 (attendus = 169, ramenes = 85.00, corrects = 10.00)
  rappel = 0.059 precision = 0.118 f-mesure = 0.079
- classe 1850 (attendus = 169, ramenes = 187.00, corrects = 17.00)
  rappel = 0.101 precision = 0.091 f-mesure = 0.096
- classe 1860 (attendus = 170, ramenes = 146.00, corrects = 13.00)
  rappel = 0.076 precision = 0.089 f-mesure = 0.082
- classe 1870 (attendus = 169, ramenes = 102.00, corrects = 10.00)
  rappel = 0.059 precision = 0.098 f-mesure = 0.074
- classe 1880 (attendus = 169, ramenes = 166.00, corrects = 19.00)
  rappel = 0.112 precision = 0.114 f-mesure = 0.113
- classe 1890 (attendus = 197, ramenes = 83.00, corrects = 7.00)
  rappel = 0.036 precision = 0.084 f-mesure = 0.050
- classe 1900 (attendus = 203, ramenes = 156.00, corrects = 15.00)
  rappel = 0.074 precision = 0.096 f-mesure = 0.084
- classe 1910 (attendus = 201, ramenes = 83.00, corrects = 10.00)
  rappel = 0.050 precision = 0.120 f-mesure = 0.070
- classe 1920 (attendus = 200, ramenes = 133.00, corrects = 15.00)
  rappel = 0.075 precision = 0.113 f-mesure = 0.090
- classe 1930 (attendus = 200, ramenes = 98.00, corrects = 10.00)
  rappel = 0.050 precision = 0.102 f-mesure = 0.067
- classe 1940 (attendus = 198, ramenes = 494.00, corrects = 82.00)
  rappel = 0.414 precision = 0.166 f-mesure = 0.237
- sur l'ensemble des 15 classes
  macro rappel = 0.116 macro precision = 0.107 macro F-mesure = 0.111
```

TABLE 10 – Fichier évalué : `equipe_3_tache_2_execution_2_1006060557.xml`

```
Evaluation du pays
- classe F (attendus = 1153, ramenes = 1313.00, corrects = 650.00)
  rappel = 0.564 precision = 0.495 f-mesure = 0.527
- classe Q (attendus = 1329, ramenes = 1169.00, corrects = 666.00)
  rappel = 0.501 precision = 0.570 f-mesure = 0.533
- sur l'ensemble des 2 classes
  macro rappel = 0.532 macro precision = 0.532 macro F-mesure = 0.532
Evaluation du journal
- classe D (attendus = 652, ramenes = 0.00, corrects = 0.00)
  rappel = 0.000 precision = 0.000 f-mesure = 0.000
- classe E (attendus = 553, ramenes = 0.00, corrects = 0.00)
  rappel = 0.000 precision = 0.000 f-mesure = 0.000
- classe M (attendus = 600, ramenes = 1313.00, corrects = 365.00)
  rappel = 0.608 precision = 0.278 f-mesure = 0.382
- classe P (attendus = 677, ramenes = 1169.00, corrects = 342.00)
  rappel = 0.505 precision = 0.293 f-mesure = 0.371
- sur l'ensemble des 4 classes
  macro rappel = 0.278 macro precision = 0.143 macro F-mesure = 0.189
```

TABLE 11 – Consolidated extra results on evaluation+reference data, top 4 each

Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	-title-text -ref -band -fft -cheb	1364	1118	54.96
2	1st	-title-text -ref -band -aggr -cheb	1363	1119	54.92
3	1st	-title-text -ref -band -aggr -diff	1362	1120	54.88
4	1st	-title-text -ref -band -fft -diff	1361	1121	54.83
Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	-title-only -ref -silence -noise -norm -aggr -eucl	1714	768	69.06
2	1st	-title-only -ref -silence -noise -norm -fft -eucl	1714	768	69.06
3	1st	-title-only -ref -low -aggr -eucl	1714	768	69.06
4	1st	-title-only -ref -noise -norm -aggr -eucl	1714	768	69.06
Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	-text-only -journal -ref -silence -noise -endp -fft -mink	1027	1455	41.38
2	1st	-text-only -journal -ref -noise -endp -aggr -mink	1027	1455	41.38
3	1st	-text-only -journal -ref -noise -endp -fft -mink	1027	1455	41.38
4	1st	-text-only -journal -ref -silence -noise -endp -aggr -mink	1027	1455	41.38
Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	-title-text -journal -ref -silence -noise -endp -aggr -eucl	1030	1452	41.50
2	1st	-title-text -journal -ref -silence -noise -endp -fft -eucl	1030	1452	41.50
3	1st	-title-text -journal -ref -silence -endp -aggr -diff	1029	1453	41.46
4	1st	-title-text -journal -ref -noise -endp -aggr -eucl	1029	1453	41.46
Run #	Guess	Configuration	GOOD	BAD	Precision,%
1	1st	-title-only -journal -ref -silence -bandstop -fft -diff	1210	1272	48.75
2	1st	-title-only -journal -ref -silence -bandstop -aggr -diff	1209	1273	48.71
3	1st	-title-only -journal -ref -bandstop -aggr -eucl	1201	1281	48.39
4	1st	-title-only -journal -ref -silence -bandstop -fft -eucl	1201	1281	48.39

4 Conclusion

We presented a MARF approach to DEFT2010 with some we believe interesting experiments and results. While the macro precision and recall leave a lot desired, we also have in indication which approaches to try and refine further. The author believes this is the only such a comprehensive study of multiple algorithm combinations and configurations used. We plan to release `DEFT2010App` as open-source at `marf.sf.net` following the documented description and the follow up analysis in (Mokhov, 2010a).

4.1 Challenges and Disadvantages of the Approach

While MARF offers a great experimentation platform, it also presents some challenges, limitations, and resulting disadvantages of using it. Here are some :

– Too many options and experiments to try (while also an advantage, it does not allow to complete all the planned experiments on time even when running on multiple machines in distributed manner).
– Unexpected UTF8 differences of compiled `.class` files on Linux vs. MacOS Java. The one on MacOS was not UTF8 by default for the strings found in class causing mismatching on the accented words such as "L'Est Républicain" and "Québec" and forcing to redo the experiments.
– Current I/O handling required files on the file system, so each article portion was serialized under its own file – a lot of I/O that hammers the performance.

4.2 Future Work / Travaux Futurs

– Add run-time statistics, recall, and f-measure to the reports.
– Complete testing the "slow" configurations mentioned in Section 2.3.
– Complete the majority of the ongoing experiments listed easrlier (to be reported in (Mokhov, 2010a)).
– MARF internally maintains a number of metrics, other than the macro precision, so we need to be able to output recall, f-measure, run-time, and other metrics in a readable way is another point to work on.
– Explore and exploit the second guess statistics.
– Explore dynamic classifier ensembles (Cavalin *et al.*, 2010).

TABLE 12 – Consolidated extra results per experiment per class on evaluation+reference data

Run #	text-title	Location	GOOD	BAD	Precision,%
1	1st	Quebec	34013	29861	53.25
2	1st	France	27657	27746	49.92
Run #	title-only	Location	GOOD	BAD	Precision,%
1	1st	Quebec	48669	16452	74.74
2	1st	France	33818	22679	59.86
Run #	text-only-journal	Journal	GOOD	BAD	Precision,%
1	1st	Le Monde	22885	6515	77.84
2	1st	L'Est Republicain	18359	8738	67.75
3	1st	Le Devoir	3989	27959	12.49
4	1st	La Presse	2861	30312	8.62
5	2nd	Le Monde	23410	5990	79.63
6	2nd	L'Est Republicain	19072	8025	70.38
7	2nd	Le Devoir	20804	11144	65.12
8	2nd	La Presse	18752	14421	56.53
Run #	text-title-journal	Journal	GOOD	BAD	Precision,%
1	1st	Le Monde	22786	6614	77.50
2	1st	L'Est Republicain	18411	8686	67.94
3	1st	Le Devoir	3880	28068	12.14
4	1st	La Presse	2974	30199	8.97
5	2nd	Le Monde	23318	6082	79.31
6	2nd	L'Est Republicain	19166	7931	70.73
7	2nd	Le Devoir	20680	11268	64.73
8	2nd	La Presse	19042	14131	57.40
Run #	title-only-journal	Journal	GOOD	BAD	Precision,%
1	1st	Le Monde	22887	6513	77.85
2	1st	L'Est Republicain	18418	8679	67.97
3	1st	Le Devoir	8686	23262	27.19
4	1st	La Presse	8324	24849	25.09
5	2nd	Le Monde	23530	5870	80.03
6	2nd	L'Est Republicain	19485	7612	71.91
7	2nd	Le Devoir	24466	7482	76.58
8	2nd	La Presse	24455	8718	73.72

TABLE 13 – Fichier évalué : `equipe_3_tache_2_execution_3-testing.sh-title-text-ref-band-fft-cheb.prexml.xml`

Evaluation du pays
- classe F (attendus = 1153, ramenes = 1521.00, corrects = 778.00) rappel = 0.675 precision = 0.512 f-mesure = 0.582
- classe Q (attendus = 1329, ramenes = 961.00, corrects = 586.00) rappel = 0.441 precision = 0.610 f-mesure = 0.512
- sur l'ensemble des 2 classes macro rappel = 0.558 macro precision = 0.561 macro F-mesure = 0.559
Evaluation du journal
- classe D (attendus = 652, ramenes = 0.00, corrects = 0.00) rappel = 0.000 precision = 0.000 f-mesure = 0.000
- classe E (attendus = 553, ramenes = 0.00, corrects = 0.00) rappel = 0.000 precision = 0.000 f-mesure = 0.000
- classe M (attendus = 600, ramenes = 1521.00, corrects = 372.00) rappel = 0.620 precision = 0.245 f-mesure = 0.351
- classe P (attendus = 677, ramenes = 961.00, corrects = 311.00) rappel = 0.459 precision = 0.324 f-mesure = 0.380
- sur l'ensemble des 4 classes macro rappel = 0.270 macro precision = 0.142 macro F-mesure = 0.186

TABLE 14 – Fichier évalué : `equipe_3_tache_2_execution_3-testing.sh-title-only-journal-ref-silence-noise-norm-aggr-eucl.prexml.xml`

Evaluation du pays
- classe F (attendus = 1153, ramenes = 1537.00, corrects = 874.00) rappel = 0.758 precision = 0.569 f-mesure = 0.650
- classe Q (attendus = 1329, ramenes = 945.00, corrects = 666.00) rappel = 0.501 precision = 0.705 f-mesure = 0.586
- sur l'ensemble des 2 classes macro rappel = 0.630 macro precision = 0.637 macro F-mesure = 0.633
Evaluation du journal
- classe D (attendus = 652, ramenes = 511.00, corrects = 178.00) rappel = 0.273 precision = 0.348 f-mesure = 0.306
- classe E (attendus = 553, ramenes = 746.00, corrects = 376.00) rappel = 0.680 precision = 0.504 f-mesure = 0.579
- classe M (attendus = 600, ramenes = 791.00, corrects = 472.00) rappel = 0.787 precision = 0.597 f-mesure = 0.679
- classe P (attendus = 677, ramenes = 434.00, corrects = 163.00) rappel = 0.241 precision = 0.376 f-mesure = 0.293
- sur l'ensemble des 4 classes macro rappel = 0.495 macro precision = 0.456 macro F-mesure = 0.475

TABLE 15 – Fichier évalué : equipe_3_tache_2_execution_3-testing.sh-text-only-journal-ref-silence-noise-endp-fft-mink.prexml.xml

Evaluation du pays
- classe F (attendus = 1153, ramenes = 2075.00, corrects = 1004.00) rappel = 0.871 precision = 0.484 f-mesure = 0.622
- classe Q (attendus = 1329, ramenes = 407.00, corrects = 258.00) rappel = 0.194 precision = 0.634 f-mesure = 0.297
- sur l'ensemble des 2 classes macro rappel = 0.532 macro precision = 0.559 macro F-mesure = 0.545
Evaluation du journal
- classe D (attendus = 652, ramenes = 318.00, corrects = 108.00) rappel = 0.166 precision = 0.340 f-mesure = 0.223
- classe E (attendus = 553, ramenes = 995.00, corrects = 410.00) rappel = 0.741 precision = 0.412 f-mesure = 0.530
- classe M (attendus = 600, ramenes = 1080.00, corrects = 476.00) rappel = 0.793 precision = 0.441 f-mesure = 0.567
- classe P (attendus = 677, ramenes = 89.00, corrects = 33.00) rappel = 0.049 precision = 0.371 f-mesure = 0.086
- sur l'ensemble des 4 classes macro rappel = 0.437 macro precision = 0.391 macro F-mesure = 0.413

TABLE 16 – Fichier évalué : equipe_3_tache_2_execution_3-testing.sh-title-text-journal-ref-silence-noise-endp-aggr-eucl.prexml.xml

Evaluation du pays
- classe F (attendus = 1153, ramenes = 2089.00, corrects = 1010.00) rappel = 0.876 precision = 0.483 f-mesure = 0.623
- classe Q (attendus = 1329, ramenes = 393.00, corrects = 250.00) rappel = 0.188 precision = 0.636 f-mesure = 0.290
- sur l'ensemble des 2 classes macro rappel = 0.532 macro precision = 0.560 macro F-mesure = 0.546
Evaluation du journal
- classe D (attendus = 652, ramenes = 279.00, corrects = 96.00) rappel = 0.147 precision = 0.344 f-mesure = 0.206
- classe E (attendus = 553, ramenes = 996.00, corrects = 414.00) rappel = 0.749 precision = 0.416 f-mesure = 0.535
- classe M (attendus = 600, ramenes = 1093.00, corrects = 480.00) rappel = 0.800 precision = 0.439 f-mesure = 0.567
- classe P (attendus = 677, ramenes = 114.00, corrects = 40.00) rappel = 0.059 precision = 0.351 f-mesure = 0.101
- sur l'ensemble des 4 classes macro rappel = 0.439 macro precision = 0.387 macro F-mesure = 0.412

TABLE 17 – Fichier évalué : equipe_3_tache_2_execution_3-testing.sh-title-only-ref-silence-noise-norm-aggr-eucl.prexml.xml

Evaluation du pays
- classe F (attendus = 1153, ramenes = 985.00, corrects = 685.00) rappel = 0.594 precision = 0.695 f-mesure = 0.641
- classe Q (attendus = 1329, ramenes = 1497.00, corrects = 1029.00) rappel = 0.774 precision = 0.687 f-mesure = 0.728
- sur l'ensemble des 2 classes macro rappel = 0.684 macro precision = 0.691 macro F-mesure = 0.688
Evaluation du journal
- classe D (attendus = 652, ramenes = 0.00, corrects = 0.00) rappel = 0.000 precision = 0.000 f-mesure = 0.000
- classe E (attendus = 553, ramenes = 0.00, corrects = 0.00) rappel = 0.000 precision = 0.000 f-mesure = 0.000
- classe M (attendus = 600, ramenes = 985.00, corrects = 468.00) rappel = 0.780 precision = 0.475 f-mesure = 0.591
- classe P (attendus = 677, ramenes = 1497.00, corrects = 514.00) rappel = 0.759 precision = 0.343 f-mesure = 0.473
- sur l'ensemble des 4 classes macro rappel = 0.385 macro precision = 0.205 macro F-mesure = 0.267

Acknowledgments / Remerciements

Nous tenons à remercier chaleureusement tous ceux et celles qui ont contribué de près ou de loin à la réalisation de ce travail, ils se reconnaîtront : Comité d'organisation : Dominic Forest, Cyril Grouin, Lyne Da Sylva ; Les distributeurs de corpora : Cedrom-SNi, ELDA, Gallica, CNRTL ; Co-créateurs origineaux du MARF : Stephen Sinclair, Ian Clément, Dimitrios Nicolacopoulos et les contributeurs du MARF R&D Group ; Drs. Leila Kosseim, Sabine Bergler, Ching Y. Suen, Joey Paquet, Mourad Debbabi ; Michelle Khalifé ; CRSNG ; La Faculté de Génie et Informatique, Université Concordia.

Références

ABDI H. (2007). Distance. In N. J. SALKIND, Ed., *Encyclopedia of Measurement and Statistics*, Thousand Oaks (CA) : Sage.

BERNSEE S. M. (1999–2005). The DFT "à pied" : Mastering the Fourier transform in one day. [online]. http://www.dspdimension.com/data/html/dftapied.html.

CAVALIN P. R., SABOURIN R. & SUEN C. Y. (2010). Dynamic selection of ensembles of classifiers using contextual information. In *Multiple Classifier Systems*, LNCS 5997, p. 145–154.

FOREST D., GROUIN C., SYLVA L. D. & DEFT (2010). Campagne DÉfi Fouille de Textes (DEFT) 2010. [online], http://www.groupes.polymtl.ca/taln2010/deft.php.

GARCIA E. (2006). Cosine similarity and term weight tutorial. [online]. http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html.

GROUIN C. (2010a). Corpus d'archives de presse (1800–1944). [online]. Source : Gallica, BNF, http://gallica.bnf.fr/.

GROUIN C. (2010b). Corpus presse francophone (1999, 2002, 2003) de La Presse, la campagne DEFT2010. [online]. Source : Cedrom-SNi, http ://www.cedrom-sni.com/.

GROUIN C. (2010c). Corpus presse francophone (1999, 2002, 2003) de Le Devoir, la campagne DEFT2010. [online]. Source : Cedrom-SNi, http://www.cedrom-sni.com/.

GROUIN C. (2010d). Corpus presse francophone (1999, 2002, 2003) de Le Monde, la campagne DEFT2010. [online]. Source : ELDA, http ://www.elda.org/.

GROUIN C. (2010e). Corpus presse francophone (1999, 2002, 2003) de L'Est Républicain, la campagne DEFT2010. [online]. Source : CNRTL, ATILF (CNRS), http://www.cnrtl.fr/corpus/estrepublicain/.

HAMMING R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, **26**(2), 147–160. See also http://en.wikipedia.org/wiki/Hamming_distance.

HARIDAS S. (2006). Generation of 2-D digital filters with variable magnitude characteristics starting from a particular type of 2-variable continued fraction expansion. Master's thesis, Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada.

IFEACHOR E. C. & JERVIS B. W. (2002). *Speech Communications*. New Jersey, USA : Prentice Hall.

JURAFSKY D. S. & MARTIN J. H. (2000). *Speech and Language Processing*. Upper Saddle River, New Jersey 07458 : Prentice-Hall, Inc., Pearson Higher Education. ISBN 0-13-095069-6.

KHALIFÉ M. (2004). Examining orthogonal concepts-based micro-classifiers and their correlations with noun-phrase coreference chains. Master's thesis, Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada.

MAHALANOBIS P. C. (1936). On the generalised distance in statistics. In *Proceedings of the National Institute of Science of India 12*, p. 49–55. Online at http://en.wikipedia.org/wiki/Mahalanobis_distance.

MARTIN J. H. (2003). The CYK probabilistic parsing algorithm. [online] ; a book insert. http://www.cs.colorado.edu/~martin/SLP/New_Pages/pg455.pdf.

MOKHOV S., CLEMENT I., SINCLAIR S. & NICOLACOPOULOS D. (2002–2003). Modular Audio Recognition Framework. Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada. Project report, http://marf.sf.net, last viewed April 2010.

MOKHOV S. A. (2008a). Choosing best algorithm combinations for speech processing tasks in machine learning using MARF. In S. BERGLER, Ed., *Proceedings of the 21st Canadian AI'08*, p. 216–221, Windsor, Ontario, Canada : Springer-Verlag, Berlin Heidelberg. LNAI 5032.

MOKHOV S. A. (2008b). Study of best algorithm combinations for speech processing tasks in machine learning using median vs. mean clusters in MARF. In B. C. DESAI, Ed., *Proceedings of C3S2E'08*, p. 29–43, Montreal, Quebec, Canada : ACM. ISBN 978-1-60558-101-9.

MOKHOV S. A. (2008–2010). WriterIdentApp – Writer Identification Application. Unpublished.

MOKHOV S. A. (2010a). Complete complimentary results report of the MARF's NLP approach to the DEFT 2010 competition. [online]. http://arxiv.org/abs/1006.3787.

MOKHOV S. A. (2010b). Cryptolysis : A Framework for Verification of Optimization Heuristics for the Automated Cryptanalysis of Classical Ciphers and Natural Language Word Segmentation. In *Proceedings of SERA 2010*, p. 295–302 : IEEE Computer Society.

MOKHOV S. A. (2010c). Evolution of MARF and its NLP framework. In *Proceedings of C3S2E'10*, p. 118–122 : ACM.

MOKHOV S. A. & DEBBABI M. (2008). File type analysis using signal processing techniques and machine learning vs. file unix utility for forensic analysis. In O. GOEBEL, S. FRINGS, D. GUENTHER, J. NEDON & D. SCHADT, Eds., *Proceedings of the IT Incident Management and IT Forensics (IMF'08)*, p. 73–85, Mannheim, Germany : GI. LNI140.

MOKHOV S. A., SINCLAIR S., CLEMENT I., NICOLACOPOULOS D. & THE MARF RESEARCH & DEVELOPMENT GROUP (2002–2010). SpeakerIdentApp – Text-Independent Speaker Identification Application. Published electronically within the MARF project, http://marf.sf.net. Last viewed February 2010.

MOKHOV S. A., SONG M. & SUEN C. Y. (2009). Writer identification using inexpensive signal processing techniques. In T. SOBH & K. ELLEITHY, Eds., *Innovations in Computing Sciences and Software Engineering ; Proceedings of CISSE'09*, p. 437–441 : Springer. ISBN : 978-90-481-9111-6, online at : http://arxiv.org/abs/0912.5502.

MOKHOV S. A. & THE MARF RESEARCH & DEVELOPMENT GROUP (2003–2010a). LangIdentApp – Language Identification Application. Published electronically within the MARF project, http://marf.sf.net. Last viewed February 2010.

MOKHOV S. A. & THE MARF RESEARCH & DEVELOPMENT GROUP (2003–2010b). ProbabilisticParsingApp – Probabilistic NLP Parsing Application. Published electronically within the MARF project, http://marf.sf.net. Last viewed February 2010.

MOKHOV S. A. & VASSEV E. (2009). Leveraging MARF for the simulation of the securing maritime borders intelligent systems challenge. In *Proceedings of the Huntsville Simulation Conference (HSC'09)* : SCS. To appear.

NIKSIC H. & FREE SOFTWARE FOUNDATION, INC. (1996–2009). wget – the non-interactive network downloader. [online]. http://www.gnu.org/software/wget/manual/wget.html, last viewed June 2010.

O'SHAUGHNESSY D. (2000). *Speech Communications*. New Jersey, USA : IEEE.

PRESS W. H. (1993). *Numerical Recipes in C*. Cambridge, UK : Cambridge University Press, second edition.

S. J. RUSSELL & P. NORVIG, Eds. (1995). *Artificial Intelligence : A Modern Approach*. New Jersey, USA : Prentice Hall. ISBN 0-13-103805-2.

THE MARF RESEARCH AND DEVELOPMENT GROUP (2002–2010). The Modular Audio Recognition Framework and its Applications. [online]. http://marf.sf.net and http://arxiv.org/abs/0905.1235, last viewed April 2010.

THE SPHINX GROUP AT CARNEGIE MELLON (2007–2010). The CMU Sphinx group open source speech recognition engines. [online]. http://cmusphinx.sourceforge.net.

VAILLANT P., NOCK R. & HENRY C. (2006). Analyse spectrale des textes : détection automatique des frontières de langue et de discours. In *Verbum ex machina : Actes de la 13eme conference annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2006)*, p. 619–629. Online at http://arxiv.org/abs/0810.1212.

μ-*Alida*: expérimentations autour de la catégorisation multi-classes basée sur *Alida*

Adil El Ghali[1] Yann Vigile Hoareau[1,2]
(1) LUTIN UserLab, Cité des Sciences, 75019 Paris
(2) Université Paris 8, 93200 Saint Denis
elghali@lutin-userlab.fr, hoareau@lutin-userlab.fr

Résumé.

Dans article nous présentons le déroulement de notre concours DEFT'10, dans lequel nous nous sommes appuyés sur l'approche *Alida*. Nous introduisons quelques unes des améliorations apportées à l'approche et les illustrons par les résultats des exécutions soumises qui avaient pour but de les tester. Pour la tâche de variation diachronique nous avons réalisé avec l'aide de nos volontaires [1] une correction du corpus pour tester les effets de corpus bruités sur nos systèmes.

Abstract.

This paper presents our work in the context of the DEFT contest. We introduce some of the enhancements to *Alida*, the approach used. And we illustrate some of them with the results of the submitted runs. In the context of the task 1, we made some correction on the original corpus in order to observe the effects of noisy data on our systems.

Mots-clés : Catégorisation de documents, Random indexing, Alida.

Keywords: Classification, Random indexing, Alida.

1 Introduction

Le thème de la sixième édition du DÉfi Fouille de Textes (DEFT'10), est l'étude des variations diachroniques et géographique du français. Deux tâches nous ont été proposées par les organisateurs. La première consistait à identifier la décennie de publication d'articles de journaux sur une période comprise entre 1800 et 1944. La deuxième à identifier l'origine géographique de chaque document (pays d'origine) dans un corpus de presse rassemblant des titres provenant de France et du Québec.

Notre travail pour cette édition du DEFT'2010, dans la lignée de l'édition 2009 (Hoareau *et al.*, 2009b), a consisté à tenter d'apporter des améliorations à notre approche cognitive de catégorisation de textes basée sur l'exploitation des espaces sémantiques : *Alida*. Nous nous sommes principalement attelé à élaborer et à tester en utilisant le corpus du DEFT'10 des méthodes permettant de combiner plusieurs instances d'Alida, pour la catégorisation multi-classes de documents.

Cette article est organisé comme suit, dans une première partie nous rappelons les fondements d'*Alida*,

1. Sylvain Baron (Bytewise), Louis-Gabriel Pouillot (Hibox) et Kaoutar El Ghali

Actes du sixième défi fouille de texte, DEFT2010, Montréal, Canada, 23 juillet 2010.
Proceedings of the Sixth DEFT Workshop, DEFT2010, Montréal, Canada, 23rd July 2010.
Pages 51-56

particulièrement sur les algorithmes d'attribution de catégories qui ont été développées et testées pour les besoins du DEFT'10. Nous présenterons ensuite les méthodes de combinaison de plusieurs instances d'*Alida*. Dans une deuxième partie, nous décrivons le déroulement de notre DEFT'10, en présentant les traitement réalisés sur le corpus et l'application de notre approche aux deux tâches. Nous conclurons notre article par une discussion des résultats et quelques perspectives de notre recherche.

2 Alida : une approche cognitive de la catégorisation de textes

Alida, l'approche issue des travaux (Hoareau *et al.*, 2009a; El Ghali *et al.*, 2009) que nous avons développé à partir de notre précédente participation au DEFT'09, se base sur une représentation des mots et des documents d'un corpus dans un espace sémantique (Karlgren & Sahlgren, 2001) implantant l'hypothèse distributionelle de (Harris, 1968). Cet espace est construit en utilisant Random Indexing (Sahlgren, 2006, 2005) et son implantation `semanticvectors` (Widdows & Ferraro, 2008).

Dans espace sémantique, sont représentés par leurs vecteurs, les documents de toutes les catégories. La phase d'apprentissage consiste à construire un vecteur prototype pour chaque catégorie en sommant l'ensemble des vecteurs de ses documents, puis à partitionner chaque catégorie en sous-catégories que nous appelons *cibles* représentant différents sous-prototypes de la catégorie.

Une fois les cibles constitués pour chaque catégorie, il s'agit de proposer des algorithmes pour attribuer la bonne catégorie à un vecteur-sonde (un document à catégoriser).

2.1 Attribution de catégories dans Alida

L'une des idées fondatrice d'*Alida*, présentée dans la section précédente, est de considérer pour chacune des catégories une décomposition en cibles. Par example, étant données, des catégories C et D, pour un nombre de cibles n, on calcule la similarité d'un document entrant d avec les cibles $C_1, ...C_n$ et $D_1, ..., D_n$. Se pose alors le problème de combiner de manière efficace ces $2 * n$ scores de similarité pour affecter la bonne catégorie à un document. Pour ce faire, nous avons élaboré et testé plusieurs méthodes de combinaison que nous décrivons brièvement ci-après :

Duel dans cette méthode, pour un document d à catégoriser, on compare deux à deux les valeurs de similarité entre d et les cibles de même rang (C_i avec D_i) et on distribue pour chaque rang i un nombre de points m entre les catégories de telle sorte que si la cible de rang i de la catégorie C : C_i est plus similaire au document à catégoriser que la cible de la catégorie D : D_i alors le nombre de points attribué, pour le document d, à la catégorie C est supérieur au nombre de points attribué á la catégorie D :

$$\mathsf{sim}(d, C_i) > \mathsf{sim}(d, D_i) \Rightarrow \mathsf{score}(d, C_i) > \mathsf{score}(d, D_i)$$

Par exemple, pour deux catégories C et D si le nombre de points à distribuer $m = 1$, si C_i est plus similaire que D_i pour un document d donné, alors $\mathsf{score}(d, C_i) = 1$ et $\mathsf{score}(d, D_i) = 0$.

Le score final d'une catégorie C pour un document d étant la somme des scores obtenu par toutes les cibles C_i de C. Et la catégorie attribuée au document d est celle qui aura obtenu le score le plus élevé.

Duel pondéré cette méthode de combinaison des similarités des cibles étend le principe du duel pour prendre en compte de manière algébrique le poids ce certaines cibles pour une catégorie donné. Cette prise en compte du poids de certaines cibles rend compte de l'importance qu'on accorde durant le processus de catégorisation par *Alida* à la particularité de certaines cibles. On peut, par exemple, favoriser les cibles qui contiennent les documents les plus typiques d'une catégorie en associant un poids élevé aux cibles de rang inférieur (les cibles les plus proches du prototype de la catégorie).

Le score final d'une catégorie C pour un document d est alors donné par la sommes des scores de ces cibles, pondéré par leurs poids respectifs :

$$\mathsf{score}(d, C) = \sum_i poids_i * \mathsf{score}(d, C_i)$$

MaxSim dans cette méthode, le score attribue à chaque catégorie C est la valeur de similarité la plus élevé de ses cibles avec le document à catégoriser. Il s'agit ici de considérer que chaque catégorie est représenté par sa cible la plus similaire au document à catégoriser :

$$\mathsf{score}(d, C) = \max_i \ \mathsf{score}(d, C_i)$$

SumSim (pondéré) dans ces méthodes, le score d'une catégorie C est obtenu en sommant les valeurs de similarité des cibles – éventuellement pondéré par le poids de leurs rangs – de C_i de C pour un document donné :

$$\mathsf{score}(d, C) = \sum_i \mathsf{sim}(d, C_i) \ \text{ resp. } \ \mathsf{score}(d, C) = \sum_i poids_i * \mathsf{sim}(d, C_i)$$

3 μ-*Alida*

Alida tout en étant conçu pour catégoriser des documents suivants plusieurs catégories, obtient de moins bonnes performances quand le nombre de ces catégories est important. Nous avons donc voulu tester si la combinaison de plusieurs instances d'*Alida* sur le même corpus en groupant des catégories dans certaines instances pouvait améliorer les performances. Deux types de combinaisons ont été envisagée :

(i) une combinaison hiérarchique, ou la catégorisation obtenu par une instance d'*Alida* ayant un nombre de catégories i ne pouvait être remise en cause par une instance d'*Alida* ayant un nombre de catégories $j > i$.

(ii) une combinaison algébrique, ou les résultats des catégorisations des différentes instances est projetté sur les suivantes. Nous avons choisi d'implanter ce type de combinaison pour le DEFT'10.

3.1 Projection algébrique

Soientt $\{C_1, ..., C_n\}$ l'ensemble des catégories d'un corpus, et $n_1, ..., n_i \in [2, \frac{n}{2}]$ des diviseurs de n. Le principe de μ-*Alida* est de créer $i + 1$ instance de *Alida*, la première correspond aux catégories d'origine, les i autres instances ayant pour catégories pour $j \in [1, i]$, $\{C'_1, ...C'_{n_j}\}$ avec C'_k la catégorie composée par l'union de $\frac{n}{n_j}$ catégories C_m successives de rang k.

Par exemple, pour $n = 4$, on aura une instance d'*Alida* avec comme catégories $\{C_1, C_2, C_3, C_4\}$ et une deuxième instance ayant pour catégories $\{C'_1, C'_2\}$ avec $C'_1 = C_1 \cup C_2$ et $C'_2 = C_3 \cup C_4$.

Chaque instance d'*Alida* permet d'associer à un document inconnu un score dans chacune des catégories (cf. 2.1). On obtient donc pour chaque document un tableau de scores de longueur n_j. Il ne nous reste plus qu'à projeter les tableaux de scores, après les avoir normalisé, les uns sur les autres en commençant par les instances ayant le plus petit nombre de catégories, comme le montre la figure 1.

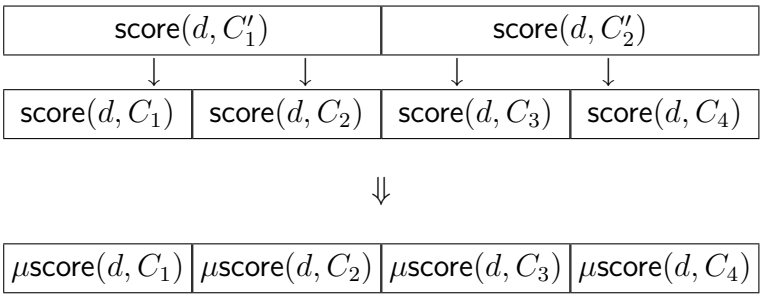


FIGURE 1 – Projection algébrique

4 Tâche 1 : Variation diachronique

4.1 Nettoyage du corpus

Le corpus diachronique qui nous a été fourni était issu de l'OCRisation de documents. il intègrait des erreurs de reconnaissance de caractères, telles que la suppression d'espace, des mots mal reconnu, ...

Nous avons décidé de tirer partie de cette particularité pour évaluer l'effet de ce genre d'erreurs sur nos algorithmes. Pour ce faire, une étape de correction partielle du corpus OCRisé à été réalisé avec l'aide précieuse de nos partenaires [2].

Plusieurs traitements ont été réalisé sur le corpus diachronique. Premièrement, une correction manuelle d'échantillons du corpus a été effectuée. Ensuite, ces corrections on été propagée sur l'ensemble du corpus. Enfin, des corrections à partir d'un correcteur orhographique opensource [3] ont été réalisés. Ces corrections sont de deux types :

1. la désagglutination : de nombreuses séquences étaient agglutinées, le passage du correcteur a permis de désagglutiner un grand nombre d'occurences.
2. le repérage d'erreurs systématiques : une modification des règles du correcteur a été nécessaire pour prendre en compte les erreurs les plus courantes de l'OCR. Par exemple, les mots contenant un caractère de ponctuation ont été traité par un ensemble de règles permettant d'obtenir des substitutions telles que : *!es* → *les*, *e)le* → *elle*, ...

2. Sylvain Baron (Bytewise) et Louis-Gabriel Pouillot (Hibox)
3. hunspell : `http://hunspell.sourceforge.net/`

4.2 Description des exécutions et Résultats

Pour la tâche de variation diachronique, le nombre de catégories était 15. Nous avons soumis trois exécutions, la première est une application d'*Alida* à 15 catégories sur le corpus brut, la deuxième est aussi une application d'*Alida* à 15 catégories sur le corpus corrigé et la troisième est l'application de μ-*Alida* à trois étages avec 3, 5 et 15 catégories sur le corpus brut.

Éxécution	Description	F-mesure	Médiane
#1	*Alida* corpus brut	0.116	
#2	μ-*Alida* corpus brut	0.156	0.181
#3	μ-*Alida* corpus corrigé	0.180	

TABLE 1 – Valeurs des F-mesures pour les exécutions de la tâche 1

La table 1 récupitule les performances des différentes exécutions soumises pour la tâche 1. Les résultats montrent que d'une part que μ-*Alida*, i.e. la combinaison de plusieurs instances d'*Alida*, améliore les performances par rapport à une application d'*Alida* avec le nombre de catégories initial. Et d'autre part, que les corrections du corpus améliorent aussi les performances.

5 Tâche 2 : Origine géographique

5.1 Description des exécutions et Résultats

Trois exécutions ont été soumise pour la tâche d'origine géographique, dans le but de tester l'effet de la méthode d'attribution des catégories dans les instances d'*Alida* sur les performances de μ-*Alida*.

La table 2 récupitule les performances des différentes exécutions soumises pour la tâche 2. Les résultats montrent que μ-*Alida* utilisant la méthode d'attribution Duel donne de meilleurs performance que μ-*Alida* avec SumSim. μ-*Alida* avec Duel pondéré donne les meilleurs performances pour la détermination du Pays tandis que μ-*Alida* avec Duel donne les meilleures performance pour la détermination du Journal.

Exécution	Description	Pays		Journal	
		F-mesure	Médiane	F-mesure	Médiane
#1	μ-*Alida* SumSim	0.762		0.424	
#2	μ-*Alida* Duel	0.798	0.792	0.446	0.462
#3	μ-*Alida* Duel pondéré	0.792		0.462	

TABLE 2 – Valeurs des F-mesures pour les exécutions de la tâche 2

6 Conclusion

Dans cette édition du DEFT'10, nous avons voulu tester un certain nombre d'optimisations de l'approche *Alida*, notamment en ce qui concerne les différentes méthodes d'attribution de catégorie. Nous avons aussi introduit un algorithme de combinaison de plusieurs instances d'*Alida*, qui permet d'améliorer les performances par rapport à une simple exécution d'*Alida*, dans le cas ou le nombre de catégories est important. Nous avons aussi eu l'occasion de tester les effets de la correction d'un corpus bruité sur les performances du système.

Remerciements

Nous tenons à remercier Sylvain Baron (Bytewise), Louis-Gabriel Pouillot (Hibox) et Kaoutar El Ghali pour leur aide précieuse sur la correction de corpus.

Références

EL GHALI A., HOAREAU Y. & EL GHALI K. (2009). The Episodic Memory Metaphor for Opinion Judgment Categorization. In *IADIS International Conference WWW/Internet (2)*, Rome.

HARRIS Z. (1968). *Mathematical Structures of Language*. New York : John Wiley and Son.

HOAREAU Y., EL GHALI A. & TIJUS C. (2009a). Detection of opinions and facts. a cognitive approach. In *Proceeding of Recent Advances in Natural Language Processing RANLP'09*, Borovets, Bulgaria.

HOAREAU Y. V., EL GHALI A. & LEGROS D. (2009b). Approche multi-traces et catégorisation de textes avec random indexing. In *Actes de l'atelier DEFT'09*, Paris.

KARLGREN J. & SAHLGREN M. (2001). From Words to Understanding. In Y. UESAKA, P. KANERVA & H. ASOH, Eds., *Foundations of Real-World Intelligence*. Stanford : CSLI Publications.

SAHLGREN M. (2005). An introduction to random indexing. *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*.

SAHLGREN M. (2006). *The Word-Space Model : Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Department of Linguistics Stockholm University.

WIDDOWS D. & FERRARO K. (2008). Semantic Vectors : A Scalable Open Source Package and Online Technology Management Application. In *Proceeding of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Actes du sixième défi fouille de texte, DEFT2010, Montréal, Canada, 23 juillet 2010.
Proceedings of the Sixth DEFT Workshop, DEFT2010, Montréal, Canada, 23rd July 2010.
Pages 57-68

TALN 2010, Montréal, 19–23 juillet 2010

Classification de textes en comparant les fréquences lexicales

Michel Généreux
Centro de Linguística da Universidade de Lisboa
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal

Résumé. Cet article fait état de travaux menés dans le cadre de la campagne DEFT 2010 concernant la classification de textes selon leur décennie ou leur origine. Nous détaillons d'abord l'approche adoptée ainsi que les ressources utilisées. On compare les fréquences des termes du lexique extrait d'un corpus d'apprentissage avec celles du lexique de référence, obtenant ainsi une liste de termes discriminants ou saillants pour chaque classe nous permettant d'attribuer un score à chaque document comme base de classification. Cette approche donne des résultats très compétitifs pour la classification selon l'origine et acceptable pour la classification diachronique. Nous utilisons aussi les lexiques de termes saillants servant de modèles pour la classification pour caractériser une classe de textes donnée.

Abstract. This article reports on work conducted under the tasks at DEFT 2010 concerning the classification of texts according to their decade or origin. First, we describe the approach and the resources used. We compare the frequencies of terms in the lexicon extracted from a training corpus with those extracted from a reference corpus, obtaining a list of discriminating or salient terms for each class allowing us to attribute a score to each document as a basis for classification. The approach gives very competitive results for the classification by origin and acceptable for the diachronic classification. We also use the glossaries of salient terms serving as models for the classification to characterize a given class of texts.

Mots-clés : Corpus comparables, Classification de textes, Analyse diachronique, Saillance, Correction orthographique.

Keywords: Comparable corpora, Classification of texts, Diachronic analysis, Salience, Spelling correction.

1 Introduction

Cette année le 6ième atelier *DÉfi Fouille de Texte (DEFT)* est consacré à la catégorisation de textes selon leur appartenance à une décennie (Tâche 1) ou selon leur origine (Tâche 2). Dans cet article nous présentons d'abord les méthodes que nous avons utilisées, en détaillant les ressources que nous avons mobilisées pour chacune de nos soumissions. Après quelques remarques sur la correction orthographique, nous faisons l'analyse, pour les deux tâches, des lexiques servant de modèles pour l'attribution d'un «score» à chaque texte, ce qui nous amène à discuter des termes les plus saillants pour chaque classe, des termes dont la variation diachronique est notable (incluant la disparition et l'apparition de termes) ainsi que des lexiques reliés au sport et à l'information. Finalement, nous présentons les résultats obtenus lors de la campagne et concluons.

2 Approche et Ressources Utilisées

Nous traitons l'ensemble des trois tâches comme un problème de classification. Notre approche est statistique mais contrairement à un bon nombre d'entre elles les modèles de classification que nous utilisons vont au-delà du sac de mots. L'idée de base est toutefois simple et a été utilisée dans des travaux sur le comportement diachronique d'expressions (Belica, 1996), ce qui s'apparente à la Tâche 1. Nous étendons et adaptons cette approche pour la classification de textes selon leur origine (Tâche 2). Dans cette approche, on génère une liste de termes saillants (i.e. des modèles pour chaque classe de textes) sur la base d'une comparaison fréquentielle entre les éléments lexicaux (1-grammes, 2-grammes et 3-grammes) d'un corpus d'apprentissage et d'un corpus de référence. Nous utilisons le *log odds ratio* (Baroni & Bernardini, 2004; Everitt, 1992) comme mesure statistique de la saillance d'un n-gramme. Le *log odds ratio* compare la fréquence d'occurrence de chaque n-gramme dans un corpus spécialisé (le corpus d'apprentissage) à sa fréquence d'occurrence dans un corpus de référence :

$$log\ odds\ ratio = ln(ad/cb) = ln(a) + ln(d) - ln(c) - ln(b)$$

où a est la fréquence du mot dans le corpus spécialisé, b est la taille du corpus spécialisé moins a, c est la fréquence du mot dans le corpus général et d est la taille du corpus général moins c. Une grande valeur de saillance positive indique une saillance forte, alors qu'une grande valeur négative indique un n-gramme sans importance pour la classe en question. Donc, à partir des corpus d'apprentissage, nous avons produit des modèles de classification pour chacune des classes des deux tâches (15 classes pour la Tâche 1 et 12 classes pour la Tâche 2). De plus, ces modèles se sub-divisaient en sous-classes, une pour chaque type de n-gramme (1, 2 et 3). Tous les textes ont été préalablement étiquetés morpho-syntaxiquement avec TreeTagger[1](Schmid, 1994), de telle sorte que chaque unité lexicale fût composée du lemme et de sa catégorie grammaticale (e.g. pomme_NOM).

Nous avons adopté une partie (75 Meg tokens) du corpus *frWAC*[2] (M. Baroni & Zanchetta., 2009) comme corpus de référence. FrWAC fait partie d'une collection de corpus de très grande taille récoltés sur Internet. Ce corpus est un bon candidat comme référence puisqu'il présente une diversification en thèmes et en genres. À titre illustratif, le tableau 1 présente le n-gramme le plus saillant pour chacun des modèles.

3 Tâche 1 : Classification selon la Décennie

Dans cette tâche, le corpus des décennies a été constitué à partir d'une «ocrisation» de journaux papiers, avec tout le bruit que cela implique, y compris au niveau des découpages de mots (e.g. dans le tableau 1, «der nier», «pro duire», etc.). Nous avons donc appliqué un pré-traitement à ce corpus visant à éliminer le plus possible les erreurs orthographiques. Nous avons d'abord construit un dictionnaire avec tous les mots de notre corpus de référence (847544 mots). Pour chaque mot du corpus de décennies, nous vérifions son orthographe de la manière suivante : s'il existe dans le dictionnaire, il est pris tel quel, sinon on essaie de le remplacer par le mot ayant la plus petite distance de Levenhstein avec un des mille mots les plus fréquents du dictionnaire. Pour limiter les temps de calcul, nous nous sommes aussi limité aux mots n'ayant pas plus ou moins de deux lettres de différences (en taille) et dont la distance de Levenhstein ne dépasse pas 2. De plus, puisque l'«ocrisation» coupait des mots en deux, nous avons remplacé toutes les paires de

[1] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

[2] `http://wacky.sslmit.unibo.it/doku.php`

Classe	1-gramme	2-gramme	3-gramme
Décennies			
Décennie 1800	numéro	avoyer être	liste du émigré
Décennie 1810	numéro	il con	celui qui avoyer
Décennie 1820	numéro	pro mettre	être ainsi que
Décennie 1830	numéro	voix nombreux	vif de curiosité
Décennie 1840	numéro	der nier	ce der nier
Décennie 1850	leur	un même	quelque un même
Décennie 1860	numéro	li liberté	être ad mettre
Décennie 1870	numéro	der nier	on nous écrire
Décennie 1880	floquet	on mander	question du scrutin
Décennie 1890	floquet	pro duire	se avoir dresser
Décennie 1900	unioniste	nom do	sortir do ce
Décennie 1910	pagnon	on mander	température se être
Décennie 1920	loucheur	pro chainer	gouvernement de Empire
Décennie 1930	reichsmark	gouverne mentir	franc par action
Décennie 1940	numéro	bri tannique	communiquer du haut
Pays			
Québec Sports	Red	Red Wings	but sur balle
Québec Informations	Irak	monsieur Landry	premier ministre Jean
France Sports	Roland-Garros	Bernard Sainz	Internationaux de France
France Informations	Irak	Viktor Tchernomyrdine	guerre en Irak
Journaux			
Le Devoir Sports	Red	coupe Stanley	but sur balle
Le Devoir Informations	Irak	monsieur Landry	premier ministre Jean
La Presse Sports	Hurricanes	série éliminatoire	but sur balle
La Presse Informations	Irak	Bernard Landry	premier ministre Jean
Le Monde Sports	Roland-Garros	Bernard Sainz	Internationaux de France
Le Monde Informations	Irak	Viktor Tchernomyrdine	guerre en Irak
L'Est Répub. Sports	L1	Christophe Mengin	Ford Focus WRC
L'Est Répub. Informations	Irak	Ehud Barak	Roissy-Charles de Gaulle

TAB. 1 – Lemmas les plus saillants pour toutes les classes (modèles)

mots consécutives absents dans le dictionnaire par leur amalgame, s'il existait dans le dictionnaire. Au final, 0.82% (8856) des mots ont été corrigé, incluant 0.04% (434 mots) qui ont été ré-assemblé après un mauvais découpage. Cette correction orthographique n'avait pour but que de limiter le nombre d'erreurs introduites par l'«ocrisation», nous n'avons donc pas produit d'évaluation de cette correction. Le tableau 2 donne un aperçu de certaines corrections effectuées. Le tableau 3 dresse un portrait de la densité lexicale

10 premiers mots corrigés	**10 premières paires de mots raccordées**
caraclère → caractère	pira teries →pirateries
servloes → services	mouil lages →mouillages
térêt → intérêt	séné galais →sénégalais
Icns → dans	Pyré nées-Orientales → Pyrénées-Orientales
émment → comment	corré lative →corrélative
cieuse → cause	extraor dinaires →extraordinaires
msp → est	sémi nariste →séminariste
guel → quel	Arbu signy →Arbusigny
dâmes → mêmes	renou velés →renouvelés
mirait → serait	scru puleuse →scrupuleuse

TAB. 2 – Corrections orthographiques automatiques pour la Tâche 1 : corpus d'apprentissage

(Bacelar do Nascimento, 2000) du corpus des décennies, en tout point comparable à la densité lexicale du corpus de référence.

Classe	**Nb Doc.**	**Types**	**Tokens**	**Verbes**	**Adverbes**	**Noms**	**Adjectifs**
Référence	131090	847544	75836891	14.2%	4.5%	25.8%	7.0%
Décennie 1800	252	5707	61785	16.1%	5.9%	21.5%	5.9%
Décennie 1810	252	5637	60907	17.0%	6.5%	21.0%	5.8%
Décennie 1820	252	5630	62221	17.1%	6.5%	20.9%	5.8%
Décennie 1830	252	5526	63683	18.4%	6.8%	20.4%	5.5%
Décennie 1840	252	5676	62410	18.0%	6.2%	21.1%	5.4%
Décennie 1850	252	5956	60752	17.4%	6.1%	21.2%	6.0%
Décennie 1860	251	5845	59494	17.4%	5.9%	21.6%	6.1%
Décennie 1870	252	5755	60150	17.6%	6.2%	21.5%	5.9%
Décennie 1880	252	6181	59575	17.6%	6.5%	21.4%	6.0%
Décennie 1890	224	6028	53265	17.9%	6.5%	21.7%	6.1%
Décennie 1900	218	6008	51398	17.4%	6.1%	22.7%	6.5%
Décennie 1910	220	5925	51433	17.1%	5.5%	22.4%	6.6%
Décennie 1920	221	6232	50634	16.9%	6.2%	22.2%	6.5%
Décennie 1930	221	6163	50601	16.4%	5.7%	22.4%	6.9%
Décennie 1940	223	5905	49703	16.5%	5.6%	22.3%	7.1%

TAB. 3 – Densités lexicales pour la Tâche 1 : corpus d'apprentissage

Avant de présenter les résultats de la classification, nous faisons quelques observations intéressantes concernant le comportement diachronique de certaines expressions. Tout d'abord, le tableau 4 montre

des termes qui présentent une forte corrélation positive ou négative entre leur degré de saillance et les années qui passent, et ce sur toute la période couverte par la Tâche 1, soit 1800-1940. Une corrélation positive indique une utilisation de plus en plus prononcée avec le temps, alors qu'une corrélation négative indique une utilisation de moins en moins prononcée. Pour donner une illustration un peu plus parlante

Corrélation positive entre 1800 et 1940	**Corrélation négative entre 1800 et 1940**
constituer_VER	ouvrage_NOM
catholique_ADJ	roi_NOM
début_NOM	prouver_VER
con_NOM	former_VER
Etats-Unis_NAM	auteur_NOM
durée_NOM	point_ADV
conférence_NOM	art_NOM
façon_NOM	jugement_NOM
durer_VER	crainte_NOM
section_NOM	reste_NOM
non_ADV :seulement_ADV	avoir_VER :point_ADV
ce_PRO :qui_PRO :concerner_VER	projet_NOM :de_PRP :loi_NOM

TAB. 4 – Corrélations positive et négative durant la période 1800-1940

de ces processus de renforcement ou d'affaiblissement de l'utilisation d'un terme, nous présentons sur un diagramme en bâtons (voir figure 1) deux des termes dont l'utilisation va croissant et deux dont l'utilisation va décroissant (voir figure 2). Ainsi, les termes *catholique* et *Etats-Unis* sont de plus en plus utilisé entre 1800 et 1940, alors que *roi* et *projet de loi* le sont de moins en moins. Nous avons poussé un peu plus cette notion de progression ou régression de l'utilisation d'un terme en examinant ceux qui sont apparût ou disparût durant cette période. Un terme «apparaît» s'il n'existe pas durant au moins la portion 1800-1820 et au plus 1800-1910 et est utilisé au moins une fois durant chaque décennie du reste de la période couverte par la Tâche 1. À l'inverse, un terme «disparaît» s'il est utilisé au moins une fois durant chaque décennie de la portion 1800-1820 et au plus 1800-1910 et disparaît durant le reste de la période couverte par la Tâche 1. Des exemples illustratifs sont montrés dans le tableau 5.

Les résultats de la classification pour la Tâche 1 sont montrés et discutés à la section 5.

4 Tâche 2 : Classification selon l'Origine

Dans cette tâche, il s'agissait de classifier des articles de journaux plus récents selon leur origine national (*France* ou *Québec*) et selon leur source de publication (*Le Monde*, *L'Est Républicain*, *Le Devoir* ou *La Presse*). Chaque article appartenait à la rubrique sportive ou d'information, et ce détail nous était fournie. Le tableau 6 nous informe d'abord sur la densité du corpus de la Tâche 2, ici encore en tout point comparable à celle du corpus de référence. Nous donnons ici encore quelques informations sur des termes intéressants du corpus. Cette fois, le tableau 7 montre les termes tirés du corpus d'apprentissage ayant une saillance élevée pour toutes les articles de sport, ce qui peut représenter un *lexique sportif*. Le tableau 8 montre les termes tirés du corpus d'apprentissage ayant une saillance élevée pour toutes les articles

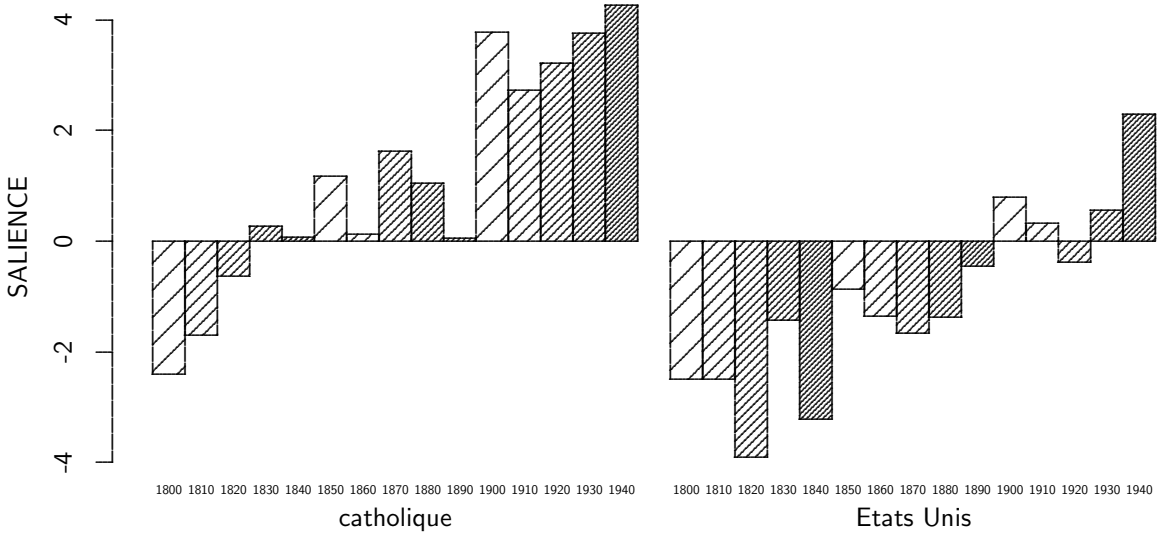


FIG. 1 – Comportement diachronique de *catholique* et *Etats-Unis*

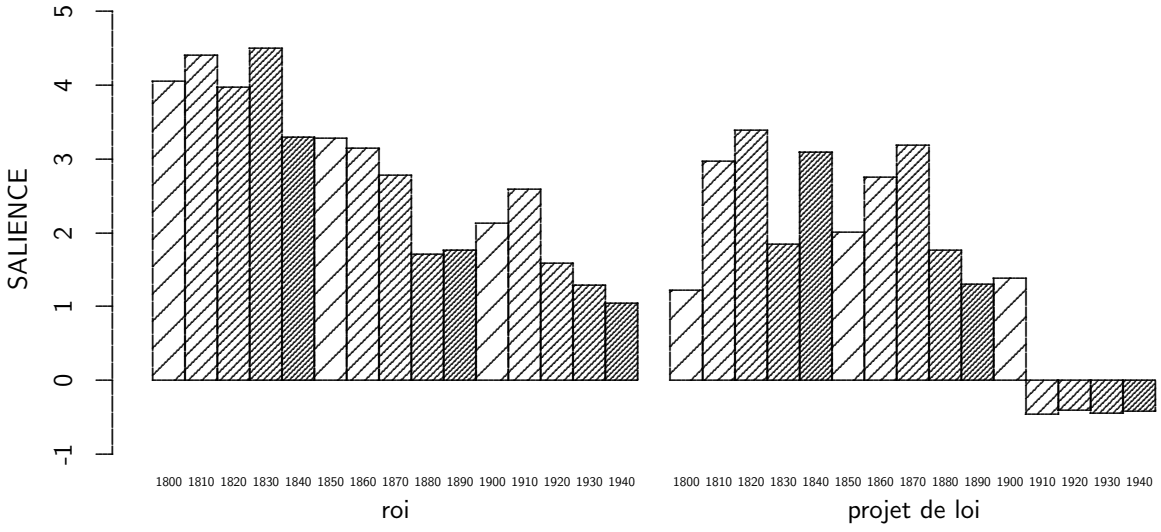


FIG. 2 – Comportement diachronique de *roi* et *projet de loi*

Apparition d'un terme durant une décennie	Disparition d'un terme durant une décennie
1830 chemin_NOM :de_PRP :fer_NOM	1830 caisse_NOM :de_PRP :amortissement_NOM
1840 clamer_VER	1830 Harpe_NAM
1840 Albert_NAM	1830 vendémiaire_NOM
1840 quelque_PRO :peu_ADV	1830 grand_ADJ :théâtre_NOM
1850 New-York_NAM	1830 partie_NOM :du_PRP :Monde_NAM
1860 tout_PRO :cas_NOM	1830 assemblée_NOM :constituant_ADJ
1870 commissariat_NOM	1830 dévoyer_VER
1870 défensif_ADJ	1830 avoyer_VER
1870 heure_NOM :actuel_ADJ	1830 étude_NOM :classique_ADJ
1880 télégramme_NOM	1830 faire_VER :le_DET :acquisition_NOM
1890 commerçant_NOM	1830 inimitié_NOM
1890 automobile_NOM	1830 division_NOM :militaire_ADJ
1900 tannique_NAM	1840 Saint-Domingue_NAM
1900 mentalité_NOM	1840 reprendre_VER :le_DET :discussion_NOM
1900 milieu_NOM :politique_ADJ	1860 avoir_VER :le_DET :malheur_NOM

TAB. 5 – Apparition et disparition d'un terme durant une décennie

Classe	Nb Doc.	Types	Tokens	Verbes	Adverbes	Noms	Adjectifs
Référence	131090	847544	75836891	14.2%	4.5%	25.8%	7.0%
Québec Sports	992	18625	418513	17.5%	6.1%	20.0%	5.1%
Québec Informations	999	22060	551252	16.2%	5.4%	22.7%	6.4%
France Sports	828	19588	356011	18.9%	7.4%	25.1%	7.0%
France Informations	900	21915	454721	15.4%	4.5%	23.2%	6.9%
Devoir Sports	475	12193	201127	17.8%	6.0%	20.1%	5.1%
Devoir Informations	501	15494	288188	16.0%	5.4%	22.6%	6.5%
Presse Sports	517	13700	217386	17.2%	6.2%	20.0%	5.1%
Presse Informations	498	15639	263064	16.4%	5.3%	22.7%	6.4%
Monde Sports	450	16956	304633	16.1%	6.3%	21.3%	6.0%
Monde Informations	450	15549	292073	15.7%	5.0%	22.6%	7.1%
Est Sports	378	11303	118575	15.4%	6.0%	20.5%	5.5%
Est Informations	450	13655	162648	14.7%	3.6%	24.2%	6.5%

TAB. 6 – Densités lexicales de la Tâche 2 : corpus d'apprentissage

Origine	1-gramme	2-gramme	3-gramme
France	Pro	Laurent Blanc	championnat de Europe
	OM	maillot jaune	quart de heure
	Nantes	club français	Coupe de Europe
	franc	Juventus Turin	Coupe de France
	Etats-Unis	Jacques Santini	million de franc
Québec	bâton	ce série	fin de semaine
	Montréal	être retirer	avoir mettre fin
	frappeur	avoir mentionner	avoir être retirer
	Robinson	six match	avoir marquer deux
	canadien	quatre coup	ne avoir donner
Devoir	s (seconde)	coupe Stanley	avoir faire savoir
	Hewitt	Lleyton Hewitt	Russe Marat Safin
	Lleyton	avoir franchir	tournoi de Wimbledon
	Seles	Monica Seles	savoir pas pourquoi
	Monica	faire savoir	Américaine Serena Williams
Presse	boulot	Brands Hatch	titan de Acadie-Bathurst
	Markov	avoir accomplir	Caroline du Nord
	Serguei	tu avoir	fin de saison
	CKAC	vendredi soir	se être emparer
	rocket	Kevin Weekes	Ligue du Champions
Monde	réputation	joueur français	ne se en
	responsabilité	cycliste international	ne pas être
	supporteurs	union cycliste	deux ou trois
	Zinedine	expliquer il	ne sembler pas
	rugby	Zinedine Zidane	union cycliste international
Est	hier	ski alpin	français du Jeux
	hier	Raymond Domenech	Ligue du Champions
	Peugeot	Britannique Tim	pays du Soleil
	10e	Justine Hénin	qui lui être
	rallye	titre olympique	dont il avoir

TAB. 7 – Lexique sportif

Origine	1-gramme	2-gramme	3-gramme
France	Blanche	Ben Laden	mise en examen
	Etats	François Hollande	mettre en examen
	Etat	parlement européen	département de Etat
	euro	tribunal correctionnel	million de euro
	Etats-Unis	maison Blanche	secrétaire de Etat
Québec	Québec	Québec avoir	attendre à ce
	État	ministre Jean	plan de action
	Montréal	gouvernement fédéral	il être aussi
	Ontario	Chrétien avoir	ce jour -ci
	Ottawa	comité exécutif	chef de accusation
Devoir	Devoir	avoir noter	conseil du ministre
	Lemieux	gouvernement québécois	plus ou moins
	caucus	se joindre	député du Bloc
	Parizeau	madame Pagé	union du municipalité
	compétition	Québec être	norme du travail
Presse	touriste	trois enfants	être encore plus
	Palestine	être signaler	nous avoir besoin
	musée	monsieur Netanyahu	dizaine de personnes
	Netanyahu	Ehud Barak	se être également
	heure	monsieur Barak	homme qui avoir
Monde	hôte	monsieur Bush	arme de destruction
	résumer	Pim Fortuyn	jouer un rôle
	Fortuyn	Viktor Tchernomyrdine	tout le monde
	Tchernomyrdine	nation uni	feuille de route
	uni	premier ministre	département de Etat
Est	hier	tribunal correctionnel	mettre en examen
	Premier	François Hollande	an avoir être
	hier	avoir requérir	million de franc
	hier	ben Laden	avoir être condamner
	correctionnel	trois homme	prendre le fuite

TAB. 8 – Lexique de l'information

d'information, ce qui peut représenter un *lexique d'information*. Les résultats de la classification pour la Tâche 2 sont montrés et discutés à la section suivante.

5 Résultats

Nous avons classifié chacun des articles du corpus de test en utilisant les ressources et la méthode décrites à la section 2. Ainsi, pour un texte donné, on compare la somme des valeurs de saillance de tous les termes présents dans les modèles. On pondère le choix des classes finales de la manière suivante : si les trois modèles «n-gramme» s'entendent[3] sur une même classe, cette classe est choisie avec un indice de confiance de un, si deux seulement s'entendent sur une classe alors celle-ci est choisie avec un indice de 0.7 et la classe unique reçoit un indice de 0.3. Finalement, si les trois classes diffèrent, alors la classe avec la somme des saillances la plus élevée reçoit un indice de 0.4 et les deux autres un indice de 0.3. Pour chaque tâche (Tâche 1, Tâche 2 - pays et Tâche 2 - journaux), nous avons produit trois soumissions. La première soumission incluait dans le calcul final les saillances de tous les termes issus des modèles produits à partir des fichiers d'apprentissage. La deuxième soumission excluait du calcul les saillances négatives et la troisième excluait du calcul les termes dits *hapax*. C'est la première soumission qui a obtenu les meilleurs résultats pour l'épreuve de classification avec les deux autres soumissions tout près derrière. Nous ne présentons ici que les détails des résultats reliés à la soumission 1. Le tableau 9 montre les résultats obtenus par l'ensemble des participants alors que les tableau 10 et 11 montrent les résultats que nous avons obtenus pour les deux tâches.

Statistique	**Tâche 1**	**Tâche 2 - Pays**	**Tâche 2 - Journaux**
Moyenne F-mesure	0.193	0.767	0.489
Médiane F-mesure	0.181	0.792	0.462
Écart-type F-mesure	0.098	0.1367	0.1887

TAB. 9 – Résultats Généraux

Pour la Tâche 2 dans son ensemble, nous faisons bonne figure, avec des F-mesure de 0.858 (Pays) et 0.630 (Journaux), comparativement à 0.767 (Pays) et 0.489 (Journaux) pour l'ensemble des participants. Les résultats sont plutôt faible pour la Tâche 1 si l'on regarde la F-mesure obtenue (0.183) mais «moyens» si l'on compare avec la F-mesure de l'ensemble des participants (0.193). Cependant, l'exactitude (0.167) reste bien au-delà de ce qu'on obtiendrait par chance (15 classes $\rightarrow$ 0.067). On remarque qu'il y a une corrélation marquée (0.53) entre le nombre de termes et le F-mesure, et une corrélation forte (0.81) entre la moyenne de la saillance et la F-mesure. L'approche est donc grandement dépendante du choix d'un corpus de référence approprié permettant de générer un nombre important de termes avec une saillance forte. Nous constatons aussi qu'il existe une forte corrélation négative entre la chronologie et la F-mesure : en d'autres termes, les résultats sont moins bons pour les décennies plus contemporaines, ce qui laisse supposer que le corpus de référence est plutôt construit à partir de documents récents, ce qui a pour conséquence de produire moins de termes saillants pour les décennies récentes. Un corpus de référence mieux réparti dans le temps permettrait sans doute d'éviter la dégradation des performances observées pour la classification des décennies plus récentes.

[3] Si la même classe obtient le score le plus élevé pour les 1-grammes, 2-grammes et 3-grammes.

Classe (Saillance moyenne, Nb de termes)	Rappel	Précision	F-mesure
1800 (8.19, 30343)	0.396	0.349	0.371
1810 (7.96, 30026)	0.181	0.212	0.195
1820 (7.95, 30387)	0.252	0.171	0.204
1830 (7.86, 31902)	0.496	0.131	0.207
1840 (7.73, 30614)	0.122	0.089	0.103
1850 (7.65, 29750)	0.211	0.134	0.164
1860 (7.59, 28722)	0.069	0.120	0.088
1870 (7.58, 29468)	0.121	0.127	0.124
1880 (7.51, 29484)	0.072	0.094	0.081
1890 (7.55, 27622)	0.130	0.137	0.133
1900 (7.73, 26901)	0.095	0.164	0.120
1910 (7.71, 26407)	0.130	0.179	0.151
1920 (7.48, 25696)	0.062	0.222	0.096
1930 (7.58, 26003)	0.053	0.227	0.086
1940 (7.66, 25078)	0.166	0.620	0.261
Exécution 1 Décennies (Exactitude = 0.167)	0.171	0.198	0.183

TAB. 10 – Résultats Exécution 1 - Tâche 1

Classe	Rappel	Précision	F-mesure
France	0.801	0.883	0.840
Québec	0.908	0.840	0.873
Exécution 1 Pays (Exactitude = 0.858)	0.854	0.861	0.858
La Presse	0.470	0.617	0.534
Le Devoir	0.598	0.543	0.569
Le Monde	0.926	0.568	0.704
L'Est Républicain	0.435	0.890	0.585
Exécution 1 Journaux (Exactitude = 0.606)	0.607	0.655	0.630

TAB. 11 – Résultats Exécution 1 - Tâche 2

6 Conclusion et Perspectives

Nous avons décrit les ressources utilisées et notre approche ainsi pour la classification de textes dans le cadre de la campagne DEFT 2010, soient une section du corpus frWAC, l'étiqueteur morpho-syntaxique TreeTagger et une méthode de classification basée sur une comparaison fréquentielle lexicale. Notons qu'il serait tout à fait possible et intéressant d'utiliser l'ensemble des textes d'apprentissage comme corpus de référence et de travailler directement sur les *lexis*, ce qui rendrait l'approche indépendante de toute ressource externe. La mesure de saillance deviendrait alors une mesure de distance sémantique entre un document à classifier et le «centre de gravité» du corpus d'apprentissage.

Cependant, l'approche décrite nous a non seulement permis d'obtenir des résultats très compétitifs pour la classification de textes selon leur origine mais aussi d'extraire des éléments lexicaux caractérisant une classe donnée ou subissant des changements diachroniques importants. Notre approche serait avantageu-

sement complétée par une contribution interdisciplinaire avec les sciences sociales (politique, histoire et sociologie) pour tirer le maximum d'information des termes saillants extraits.

Notre approche a obtenu des résultats moyens pour la classification selon la décennie, ce que nous expliquons par la faible variation diachronique du corpus de référence avec comme résultat un nombre moins important de termes saillants pour les périodes récentes. Néanmoins, cette approche comparative fût très productive et a produit des résultats intéressants, nous avons donc l'intention de l'appliquer à des textes d'autres périodes et provenant d'autres sources.

Références

BACELAR DO NASCIMENTO M. F. (2000). *Corpus de Référence du Portugais Contemporain*, In *Corpus, Méthodologie et Applications Linguistiques*, p. 25–30. Presses Univ. de Perpignan. Editor : M. Bilger.

BARONI M. & BERNARDINI S. (2004). Bootcat : Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, p. 1313–1316.

BELICA C. (1996). Analysis of temporal changes in corpora. *International Journal of Corpus Linguistics*, **1**(1), 61–73.

EVERITT B. (1992). *The analysis of contingency tables*. London : Chapman and Hall.

M. BARONI, S. BERNARDINI A. F. & ZANCHETTA. E. (2009). The wacky wide web : A collection of very large linguistically processed web-crawled corpora. In *Language Resources and Evaluation*, volume 43, p. 209–226.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

Système du LIA pour la campagne DEFT'10 : datation et localisation d'articles de presse francophones

Stanislas Oger[1] Mickael Rouvier[1] Nathalie Camelin[1] Rémy Kessler[1]
Fabrice Lefèvre[1] Juan-Manuel Torres-Moreno[1,2]
(1) Laboratoire Informatique d'Avignon, BP 91228, F-84911 Avignon, France
(2) École Polytechnique de Montréal, CP 6079, Montréal (Québec) H3C3A7 Canada
{nathalie.camelin, remy.kessler, stanislas.oger, mickael.rouvier, fabrice.lefevre,
juan-manuel.torres}@univ-avignon.fr

Résumé. Nous présentons dans cet article les systèmes développés au LIA pour la campagne d'évaluation DEFT'10. La campagne comporte deux tâches (ou *pistes*) distinctes : la première consiste à identifier la décennie de publication d'articles francophones entre 1800 et 1940 et la seconde à identifier le pays (France ou Quebec) et le journal dans lequel a été publié l'article parmi 4. Plusieurs systèmes basés sur des modèles probabilistes ont été développés pour chacune des tâches. Puis ces systèmes ont été fusionnés pour fournir une distribution globale sur l'ensemble des hypothèses, permettant une décision globale. La bonne robustesse des systèmes individuels et de leur fusion entre le corpus d'apprentissage et de test nous a permis d'obtenir de bons résultats, b ien que très contrastés selon les tâches.

Abstract. This paper describes the systems developed at LIA for the DEFT'10 evaluation campaign. This campaign includes two different tasks : (a) identifying the decade of publication of French-written articles, and (b) identifying the country where the articles were published and the journal. Several systems, all based on probabilistic models, were developed. A final fusion step provides an overall distribution on the hypotheses, allowing a better final decision. The good robustness of the individual systems and the fusion system between the training and testing corpora allowed us to obtain good results, although well contrasted over the various tasks.

Mots-clés : Méthodes probabilistes, Apprentissage automatique, Classification de textes par leur contenu, Défi DEFT.

Keywords: Stocahstic approaches, Machine learning, Text classification, DEFT challenge.

1 Introduction

La sixième édition de la campagne d'évaluation DEFT (DÉfi Fouille de Textes) a eu lieu au printemps 2010. Cette année encore un défi original en fouille de textes a été proposé à la communauté francophone avec pour objet deux tâches distinctes : les variations diachroniques et l'origine géographique de corpus de presse francophones.

La thématique Langage du Laboratoire d'Informatique d'Avignon (LIA) [1] a relevé ce défi pour la quatrième fois. Six participants se sont mobilisés pour l'occasion, avec pour la moitié d'entre eux une première participation. Lors des éditions précédentes, le LIA a toujours participé avec succès à ce défi en proposant différents systèmes basés sur des méthodes statistiques ainsi que des méthodes de fusion s'appuyant sur ces systèmes (El-Bèze *et al.*, 2005; Torres-Moreno *et al.*, 2007; El-Bèze *et al.*, 2007; Béchet *et al.*, 2008; Charton *et al.*, 2008; Torres-Moreno *et al.*, 2009).

Le défi actuel implique de classer des articles de presse francophone d'une part selon la période à laquelle ils ont été écrits et d'autre part selon le pays dans lequel ils ont été publiés. Ces deux tâches semblent au premier abord assez simples. En effet, chaque article suit un style particulier. Ce style dépend de l'auteur, certes mais cet auteur évolue à une période donnée et dans une localisation donnée. Chacun de ces paramètres induit d'une part des choix lexicaux spécifiques mais également des tournures stylistiques particulières. C'est ce style qu'il est ici question de définir. Plusieurs mots définissent une localisation plutôt qu'une autre (*e.g.* aiguisoir en fran cais québécois pour taille-crayon) mais qu'en est-il de la distinction de journaux d'un même pays ? En ce qui concerne les humains, il semble raisonnable de penser que des experts littéraires seraient capables de faire le distinguo. Qu'en est-il des méthodes d'apprentissage automatique ? D'autre part, tout francophone èducè est capable de distinguer les styles littéraires marquès (comme entre Victor Hugo et Molière, par exemple). Toutefois cette assertion est beaucoup plus sujette a caution dans le cas de deux grands écrivains de la même époque, ou d'époques proches. Cela nécessiterait certainement des compétences trés particulieres. La tâche de classification n'est donc pas aisée du tout. D'une part, les classes proposées ne sont pas si évidemment dissemblables (*e.g.* 1810 et 1820 ou « Le devoir » et « La presse »). Et d'autre part, le manque de corpus est encore et toujours un problème. En effet, les méthodes que nous proposons sont majoritairement des méthodes discriminantes basées sur des apprentissages supervisés et donc le manque de corpus annoté reste problématique. Les applications d'une telle tâche sont diverses. Elles vont du filtrage de grands corpus pour faciliter la recherche d'information ou la veille scientifique et économique jusqu'à la classification par le type de texte pour adapter les traitements linguistiques aux particularités d'un corpus.

Nous décrivons dans cet article les techniques et méthodes automatiques utilisées pour relever ce défi. La section 2 présente les tâches et les corpus de DEFT'2010. La section 3 décrit chacun des systèmes initiaux développés par les participants du LIA. Ces systèmes initiaux (niveau 1) sont ensuite utilisés par des systèmes de second niveau selon différents principes de fusion, définis dans la section 4. L'evaluation et les résultats des expériences sont rapportés et discutés en sections 5 et 6.

2 Présentation des tâches et des corpus

2.1 Variations diachroniques

Cette piste, relative à la variation diachronique, concerne l'identification de la décennie de publication d'extraits d'articles de presse français. La période couverte va de 1800 à 1944. Il s'agit donc d'une classification de texte de type uni-label sur un ensemble de 14 étiquettes. Une difficulté posée par cette classification concerne l'aspect continu (et non discret) des étiquettes : un texte écrit le 31 décembre 1849 n'aura pas la même étiquette que celui écrit le lendemain. Donc s'il peut sembler à peu prés facile, pour

1. http://www.lia.univ-avignon.fr

un humain, de classer des articles dans deux classes représentant des dates suffisamment éloignées, ne serait-ce que par le sujet évoqué. En revanche, quels sont les indices permettant de définir si un article a été écrit plutôt en 1949 qu'en 1950 ?

Le corpus d'apprentissage se compose de 3594 extraits de longueur 300 mots d'articles de quatre titres de journaux différents : « Le journal de l'empire », « Le Journal des Débats politiques et littéraires », « La Croix » et « Le Journal des Débats ». 15 décennies sont à identifier : 1800, 1810, 1820,...1940. Si un article a été écrit en 1809 alors il appartient à la décennie 1800, s'il est écrit en 1810 alors il appartient à la décennie 1810. Le tableau 1 contient le nombre d'articles pour chaque décennie du corpus d'entraînement.

Décennie	1800	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940
Nombre	252	252	252	252	252	252	251	252	252	224	218	220	221	221	223

TABLE 1 – Répartition des extraits du corpus d'apprentissage Décennie

De manière à éprouver la robustesse des systèmes, le corpus de test intègre des extraits provenant des quatre titres présents dans le corpus d'entraînement, plus un cinquième absent de ce dernier. En pratique une difficulté majeure réside dans la lisibilité du texte. En effet, ces articles ont été obtenus à la suite d'un traitement automatique de reconnaissance de caractères sur des images scannées des journeaux et souffrent d'un nombre très élevé d'erreurs de reconnaissance.

2.2 Origine géographique

L'identification de l'origine géographique de chaque document (pays d'origine) constitue la seconde piste de cette campagne. Elle repose sur des corpus de presse rassemblant deux titres provenant de France et deux autres du Québec. Le corpus d'apprentissage est composé de 3719 extraits : 60% des corpus d'origine, les 40% restants sont utilisés pour le test.

Sous-tâche 1 : Quel Pays ? Les deux pays à retrouver sont la France et le Québec. Le tableau 2 donne le nombre d'extraits de chaque pays présents dans le corpus d'apprentissage.

Pays	Québec (Q)	France (F)
Nombre	1991	1728

TABLE 2 – Répartition des étiquettes Pays sur le corpus d'apprentissage Origine

Sous-tâche 2 : Quel Journal ? Cette sous-tâche consiste à identifier de quel journal précisement provient l'extrait. Les quatre journaux sont « Le Monde » (étiquette M) et « l'Est républicain » (E) pour la France, « Le Devoir » (D) et « La Presse » (P) pour le Québec. Le tableau 3 donne le nombre d'extraits de chaque journal présents dans le corpus d'apprentissage.

Pays	Le Monde (M)	l'Est républicain (E)	Le Devoir (D)	La Presse (P)
Nombre	900	828	976	1015

TABLE 3 – Répartition des étiquettes Journal sur le corpus d'apprentissage Origine

3 Présentation des systèmes initiaux

Comme les années précédentes, chaque participant du LIA propose un ou plusieurs systèmes. Ces systèmes sont donc différents par la méthode de classification employée ou par les techniques de représentation des articles propres à l'appréhension personnelle de la tâche par chaque concepteur. Il s'agit donc conjointement d'optimiser chacun des systèmes initiaux mais aussi d'obtenir un ensemble aussi hétérogène que possible. Ce qui sera mis à profit par l'utilisation dans un second temps de techniques de fusion permettant d'obtenir le meilleur système global possible. Les techniques de fusion sont décrites dans la section 4.

3.1 Protocole expérimental

Nous présentons d'abord brièvement les choix communs. Puis, les sous-sections suivantes décrivent les différentes méthodes ainsi que chacun des systèmes de niveau 1.

Etiquettes morpho-syntaxique ou lemmatisation Lors de nos participations précédentes, l'utilisation des POS (Part-of-Speech) et lemmes était systématiquement proposé à chacun de nos systèmes. Cette année en revanche, la trop grande quantité de bruit dans les corpus due à l'OCRisation mais également à des problèmes d'encodage n'a pas permis d'effectuer une analyse syntaxique efficace des textes. Ainsi, l'utilisation des POS, lemmes ou stemmes a été abandonnée.

Validation croisée Afin de tester nos méthodes, de régler leurs paramètres et de palier au phénomène de sur-apprentissage, nous avons décidé de scinder l'ensemble d'apprentissage (A) de chaque corpus en 5 sous-ensembles approximativement de la même taille (en nombre d'articles à traiter). La procédure d'apprentissage a été la suivante : 4 des 5 sous-ensembles sont concaténés pour produire un corpus d'entraînement et le cinquième est utilisé pour le test. La procédure est effectuée cinq fois afin que chacun des sous-ensembles du corpus d'apprentissage soit utilisé une fois pour le test. Les ensembles ainsi concaténés seront appelés dorénavant ensembles de développement (D) et le restant ensemble de validation (V).

3.2 Systèmes extra-linguistique : *Mick_MLP* et *Mick_SVM*

3.2.1 Sac de mots

L'une des difficultés majeures de la tâche décennie réside dans la lisibilité du texte. En effet, les articles ont été obtenus à la suite d'un traitement OCR et souffrent énormément des erreurs de reconnaissances. Ces erreurs de reconnaissances introduisent un bruit dans les documents et peuvent dégrader les performances des classifieurs. Nous proposons une méthode de correction des erreurs de mots et une méthode de classification de texte basé sur les mots-outils.

Correction des closures Les erreurs de reconnaissance issues d'un OCR sont multiple et peuvent être regroupées en 2 catégories : les erreurs de non-mots (*non-word errors*) et les erreurs de mots-réels (*real-word errors*). Une erreur de non-mot apparaît quand un mot est interprété comme une chaîne de caractère qui n'appartient pas à la langue française. Une erreur de mot-réel apparaît quand un mot est interprété comme une chaîne de caractère qui appartient à la langue française, mais n'est pas identique à la source du

texte. Par exemple, si la phrase source « le président de la république » est reconnu par un OCR comme « la président de la répablique » , *répablique* est une erreur de non-mot et *la* est une erreur de mot-réel.

Nous nous sommes intéressés ici aux erreurs de non-mot et plus particulièrement aux troncatures des mots. Par exemple le mot « président » peut être retranscrit après l'OCR comme « pré sident » . Ces erreurs introduisent du bruit dans les classifieurs et peuvent être corrigées très facilement à l'aide d'un dictionnaire. Pour chaque document, nous extrayons les bi-grammes de mots. Pour chacun de ces bigrammes, si la concaténation des mots le composant appartient au dictionnaire alors nous remplaçons le bigramme par le mot du dictionnaire. Le dictionnaire utilisé dans notre système a été constitué à partir des articles du journal Le Monde 1987-2003 ainsi que sur celui d'ESTER (Gravier *et al.*, 2004).

Classification sur les mots-outils La plupart des méthodes de classifications de textes sont basées sur un TF-IDF (*Term Frequency-Inverse Document Frequency*) et/ou un pré-traitement linguistiques (POS, lemmatisation). Ces techniques peuvent montrer des faiblesses lorsque les textes sont trop bruités (sortie d'un OCR, sortie de transcription, etc). (Stanislas Oger, 2010) a montré, dans le cadre de la classification de genre vidéo, que l'utilisation de la fréquence d'apparition des mots outils de la langue améliore nettement les résultats par rapport à l'utilisation de TF-IDF lorsque les données textuelles sont bruitées.

3.2.2 Extraction des entités nommées

A l'aide d'une encyclopédie numérique, les entités nommées d'un document (noms de personne, lieux, etc), peuvent nous donner des informations intéressantes sur sa date de rédaction. La détection d'entités nommées sur un corpus bruité n'est pas une chose facile (contexte bruité, noms propres mal orthographiés, etc). Nous proposons ici d'utiliser le systeme LIANE. Il permet de détecter les entités nommées sur des sorties bruités comme des sorties de transcription automatique de la parole ou issues d'un processus d'OCR.

Nous proposons donc un système à trois niveaux. Le 1er niveau va détecter les entités nommés. Le 2ième niveau va vérifier que l'entité nommée est bien écrite grâce à l'outil en ligne de suggestion d'orthographe de Google. Finalement, le 3ième niveau va rechercher sur une encyclopédie numérique (Wikipédia dans notre cas) les dates correspondant aux entités nommées détéctées. Un vecteur contenant toutes les dates est ainsi créé. Chaque indice du vecteur correspond au nombre de fois où la date a été vue sur Wikipédia. Ce vecteur est ainsi ajouté au vecteur d'observation du classifieur.

3.2.3 Extraction des caractères de ponctuations

Le taux d'erreur d'un document est dû en grande partie à la qualité du document numérisé. Les OCR sont assez sensibles à la qualité du papier, à la police de caractères, à l'encre utilisée, etc. On peut donc penser que le nombre d'erreurs rencontrées est lié à la date d'écriture du document. Au plus un document est ancien, au plus le taux d'erreur est élevé. Deux critères sont traditionnellement utilisés pour modéliser le taux d'erreur d'OCR : le nombre de mot hors vocabulaire et la perplexité du document selon un modèle de langage appris sur un corpus sans erreurs. Dans ce travail, nous proposons d'utiliser la ponctuation car les artefacts d'un document (tâches d'encre, etc.) sont souvent transformés par l'OCR en signes de ponctuation. Pour chaque document nous créons donc un vecteur contenant les fréquences d'apparition des signes de ponctuation. Ce vecteur est aussi ajouté au vecteur d'observation du classifieur.

3.2.4 Classifieur

Au total, le vecteur d'observation du classifieur est constitué des mots outils (soit environ 20 000 entrées), des fréquences des signes de ponctuation (13 entrées) et des entités nommées (15 entrées). Nous avons testé, sur le corpus de test, 2 classifieurs : les machines à vecteurs supports (*Support Vector Machine*, SVM) et les réseaux de neurones de type perceptron multi-couche (MLP). Nous avons choisi pour le SVM d'utiliser un noyau linéaire et pour le MLP une topologie à 3 couches (avec une couche cachée de 1000 neurones). Les performances globales des 2 classifieurs sont comparables. Par contre les réponses données par les classifieurs sont très différentes et donc nous pouvons espérer qu'en combinant ces 2 classifieurs dans un systèmes de fusion cela puisse améliorer les résultats.

3.3 Modèle de langage à base de n-grammes de caractères : *Jmt* et *Jmt_basic*

Dans le contexte du défi DEFT'10, nous voulions savoir si les n-grammes de caractères permettaient de discriminer convenablement la classe des documents. Dans le cas affirmatif, cela présenterait plusieurs avantages par rapport aux n-grammes de mots. D'abord l'ensemble des n-grammes de caractères est considérablement plus petit que l'ensemble des n-grammes de mots. Dans l'utilisation des modèles n-grammes de caractères, on peut se passer des techniques de lissage ou de *Back-Off* (Manning & Shütze, 2000), car à la différence des mots, la plupart des caractères rencontrés dans les phases de test ont été observés. Enfin, bien que l'utilisation des caractères reste relativement différente entre les langues, la plupart des signes sont les mêmes entre les langues d'origine latine. Nous avons développé un classifieur classique incorporant des techniques élémentaires de n-grammes, mais en utilisant les caractères à la place des mots. Ces techniques, inspirées directement de l'approche probabiliste (Manning & Shütze, 2000) appliquées à la classification de texte, ont prouvé leur efficacité dans les défis DEFT précédents (El-Bèze *et al.*, 2005; Torres-Moreno *et al.*, 2007; El-Bèze *et al.*, 2007; Béchet *et al.*, 2008; Charton *et al.*, 2008). Pour une tâche de classification, on peut construire les modèles n-grammes associés aux classes recherchées, par exemple dans la tâche Origine, Pays et Journal $g \in \{P, J\}$. Le score du genre $\tilde{g}$ étant donné un document et une séquence de caracteres s, aurait pu être calculé selon le théorème de Bayes :

$$\tilde{g} = \arg\max_g P(g|s) = \arg\max_g \frac{P(s|g)P(g)}{P(s)} = \arg\max_g P(s|g)P(g) \quad (1)$$

$$\tilde{g} \approx \arg\max_g P(s|g) \approx \arg\max_g \prod_i P_g(s_i|s_{i-2}, s_{i-1}) \quad (2)$$

combinée avec une interpolation simple. Cependant, nous avons voulu en particulier étudier les algorithmes originelement conçus pour l'identification de la langue (Cavnar & Trenkle, 1994). Pour préserver toute son efficacité, nous n'avons réalisé aucun filtrage de signes de ponctuation des corpus d'apprentissage pour ce système [2]. L'algorithme proposé opère en 2 phases : (a1 la création du modèle de langage (ML) pour chaque catégorie i de document, $i = 1, \cdots, c$ et (2) le calcul de distance d'un document inconnu par rapport aux ML_i :

Phase 1. Modèles de langage $\mathbf{M}_i$

1. Découper le texte en *tokens* (chaînes de caractères séparées seulement par des espaces).
2. Génération de tous les n-grammes possibles, pour $n = 1, \cdots, 5$.

2. Nous remercions Marc El-Bèze (LIA) pour ses scripts de conversion de caractères.

3. Créer une table triée inversée pour comptabiliser les occurrences des n-grammes.

Phase 2. Calcul de distance sur un document inconnu

1. Créer un ML_x du document inconnu x au même titre que dans la phase (1).
2. Au moyen des i modèles de langage ML_i, calculer une statistique simple du rang au moyen d'une mesure de distance.
3. Initialiser un score à 0. Pour chaque n-gramme $\in ML_x$, cette distance détermine dans quelle mesure la position d'un n-gramme dans ML_x est dans la même position dans chaque ML_i. Même position, score += 0, position différente : score += | distance entre la position du n-gramme(ML_x) et n-gramme(ML_i) |. Les n-grammes $\in ML_x$ qui ne sont pas présents dans un modèle ML_i seront pénalisés par une grande valeur fixée empiriquement.
4. Le document inconnu x est attribué à la classe i avec le score le plus bas (i.e. la distance la plus petite) par rapport au modèle ML_i.

Nous avons fixé la distance maximale pour les n-grammes inconnus égale au nombre de n-grammes générés lors de la phase de création du ML. Nous avons appliqué ce modèle appéle n-grammes « basique » au corpus d'apprentissage de la tâche Origine, sans faire d'autres traitements particuliers. Pour faire varier un peu la stratégie, nous avons modifié le modèle de n-grammes précédent pour savoir si la contrainte de considérer les n-grammes de caractères sur la longueur du texte et pas sur celle de mots apportaient des éléments discriminants. Ainsi les ML ont été générés avec des n-grammes construits sur le texte vu comme une chaîne complète.

3.4 Un BoosTexter tout simplement : *Boost_basic*

Algorithme de *boosting* Le but de cet algorithme est d'améliorer la précision des règles de classification en combinant plusieurs hypothèses dites *faibles* ou peu précises.

Une hypothèse faible est obtenue à chaque itération de l'algorithme de *boosting* qui travaille en re-pondérant de façon répétitive les exemples dans le jeu d'entraînement et en ré-exécutant l'algorithme d'apprentissage précisément sur ces données re-pondérées. Cela permet au système d'apprentissage faible de se concentrer sur les exemples les plus compliqués (ou problématiques).

L'algorithme de *boosting* obtient ainsi un ensemble d'hypothèses faibles qui sont ensuite combinées en une seule règle de classification qui est un vote pondéré des hypothèses faibles et qui permet d'obtenir un score final pour chaque constituant de la liste des concepts.

Les composants du vecteur d'entrée sont passés selon la technique du sac de mots et les éléments choisis par les classifieurs simples sont alors des n-grammes sur ces composants.

Boost_basic Ce système utilise le classifieur à large marge *BoosTexter (Schapire & Singer, 2000)* basé sur l'algorithme de *boosting Adaboost* (Freund & Schapire, 1996). Il est paramétré pour faire 600 ré-exécution au maximum et à chaque exécution choisir un motif de 1 à 3 mots consécutifs.

Les paramètres d'entrée de ce système sont les formes lexicales de chaque document. Les champs <texte> et <titre> pour la tâche *Origine* et le champ <texte> pour la tâche *Décennies*. Ces entrées lexicales subissent simplement deux traitements simples.

Dans un premier temps, les références aux entités XML sont remplacées par le caractère correspondant (*e.g.* &apos devient '). Dans un deuxième temps, le texte est nettoyé de toute ponctuation (! ?., ;) et autres caractères spéciaux (()#'&), chacun de ces caractères étant remplacé par une balise lexicale tenant compte de la présence ou non d'un espace autour du caractère. Par exemple, "suivant :" deviendra "suivant BAL-2POINTS-G" et "suivant :" deviendra "suivant BAL-2POINTS" tandis que " :encore" deviendra "BAL-2POINTS-D encore". Ainsi, la nouvelle phrase obtenue ne contient que des caractères aA-zZ et ne perd pas l'information de l'utilisation du caractère espace ' '.

3.5 Une autre approche du *boosting* : *Rk_icsiboost*

Pour ce système nous avons aussi utilisé un classifieur à large marge de type Adaboost spécialisé dans le traitement de données textuelles. L'outil ICSIBOOST [3] a été utilisé. Les paramètres d'entrée du système sont les entrées lexicales de chaque document, les champs <texte> et <titre> pour la tâche Origine et le champ <texte> pour la tâche Décennie. Nous avons choisi une représentation sous forme de n-grammes de mots, celle-ci ayant obtenu les scores les plus élevés lors des nos premiers tests. Afin d'améliorer les résultats sur la tâche Décennie, nous avons remplacé l'ensemble de la ponctuation par une balise lexicale, ainsi « , » devient « _VIR_ », « . » devient « _POINT_ », etc. Nous avons tenté par ailleurs certains prétraitements linguistiques classiques (Manning & Shütze, 2000) tels que le filtrage ou la racinisation sans amélioration notable des résultats. Concernant le nombre de tours de l'algorithme (paramètre T), les expériences ont montré qu'un optimum était atteint aux alentours des 1000 tours, ceci permettant des temps d'apprentissage assez court.

3.6 Système hybride pour résoudre la tâche Décennie

Confusion entre les classes Une des difficultés évidentes de la résolution de la tâche Décennie est le caractère continu des étiquettes. En effet, l'étude des matrices de confusion des systèmes sur le corpus de développement a mis en évidence que la majorité des erreurs étaient faites à une ou deux décennies près de celle de référence. A titre d'illustration, le tableau 4 donne la matrice de confusion du système *Mick_MLP* sur la partie (A.1) du corpus d'apprentissage.

Boost_basic_3classes Afin de palier ce problème, plusieurs systèmes initiaux ont été implémentés avec pour chacun le classement d'uniquement trois décennies consécutives. Ainsi, par exemple, le classifieur dont l'étiquette médiane est 1810 n'est appris que sur les exemples du corpus d'apprentissage étiquetés 1800 ou 1810 ou 1820. Ces classifieurs ont été implémentés selon le même protocole que le système initial Boost_basic. Les résultats de ces systèmes évalués en milieu fermé (c'est-à-dire sur des exemples du corpus de validation étiquetés selon le même sous-ensemble d'étiquettes que celui de développement) sont rapportés dans le tableau 5.

Les résultats obtenus sont meilleurs que ceux du même système appris sur les 14 classes (F-mesure moyenne d'environ 26%). Il est également intéressant de noter que les scores les plus élevés sont obtenus aux deux extrémités de l'intervalle de temps 1810 et 1930 avec des F-mesures dépassant les 50%.

3. http://code.google.com/p/icsiboost/

	1800	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940
1800	28	13	2		1										
1810	4	24	6	4											
1820	2	29	14	7		1	1		2					1	
1830		3	4	22	8	2	2	2	1	1	1		1		
1840		2	1	15	11	6	3	11	6	2	1				
1850	1	7	2	10	3	7	7	7	5		1		1		
1860		4	1	7	3	4	6	4	3		3	2	2		1
1870		2		4	4	1	2	20	11	3	2	3	1		1
1880		1		5	1	1	5	7	13	6	4		2		3
1890		3	1	1	2	2	1	9	10	4	7	1	4	2	
1900	1	2	1	3	1	1	1	2	9	3	9	5	6		1
1910	1	1		1		1	1	2	5	4	4	11	6	4	13
1920			1	3			3	3	9	3	6	4	1	6	7
1930		1		2		1	1	1	5		4	2	7	11	12
1940		1		1				2	1		2			7	28

TABLE 4 – Matrice de confusion du système *Mick_MLP* sur la partie (A.1) du corpus d'apprentissage. Les colonnes correspondent aux décennies de référence tandis que les lignes correspondent à la décennie choisie par le système (score de confiance le plus élevé).

Etiquette médiane	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930
F-mesure moyenne	0.52	0.55	0.52	0.47	0.40	0.35	0.45	0.44	0.36	0.37	0.47	0.48	0.51

TABLE 5 – Performance de chacun des sous-classifieurs appris et testés sur 3 décennies consécutives.

Le choix de ne pas créer de classifieurs bi-classes pour les décennies 1800 et 1940 est également basé sur l'observation de bons résultats obtenus avec un classifieur 14-classes sur ces classes extrêmes.

Hybride_Boost_MLP Un système hybride a été élaboré à partir des systèmes *Mick_MLP* et *Boost_basic_3classes*. Le principe est le suivant :

1. la décision de *Mick_MLP* détermine la classe médiane ;
2. le système *Boost_basic_3classes* correspondant à cette classe est appliqué sur l'exemple ;
3. l'étiquette finale est celle obtenant le plus haut score de confiance pour *Boost_basic_3classes*.

Au final, les scores de confiance de l'ensemble des étiquettes est mis à 0 sauf ceux ayant obtenus un score dans *Boost_basic_3classes* où leur score est conservé.

4 Systèmes de fusion

Lors de l'évaluation chaque participant propose 1 à n systèmes et trois soumissions sont faites : meilleur classifieur pour la tâche considérée, vote majoritaire et fusion probabiliste, décrits dans les sous-sections suivantes.

4.1 Fusion majoritaire

Une des manières les plus simples de prendre en compte l'ensemble des décisions est d'effectuer un vote majoritaire. Le système de fusion majoritaire prend en considération les systèmes suivants :
– Tâche Décennie : *Boost_basic*, *Hybride_Boost_MLP*, *Mick_MLP*, *Mick_SVM* et *Rk_icsiboost* ;

– Tâche Origine : *Jmt*, *Jmt_basic*, *Boost_basic*, *Mick_SVM* et *Rk_icsiboost*.

Le vote de chaque système correspond à l'étiquette ayant obtenu le score de confiance le plus élevé. Une voix est accordée à chacun des systèmes de base définis pour la tâche donnée. Par ailleurs, un classement des systèmes selon leur performance sur le corpus d'apprentissage est établi pour chaque tâche. Ainsi, en cas d'égalité de voix sur deux ou plusieurs étiquettes, l'étiquette choisie parmi ces dernières sera celle élue par le système le plus performant.

Chaque étiquette en lice (choisie par au moins un des systèmes de base) se voit attribué un score de confiance. Ce score correspond au quotient du nombre de voix obtenues sur le nombre de votants. En cas d'égalité, l'étiquette choisie grâce au classement des performances des classifieurs est augmentée d'une voix, ainsi que le nombre de votants. Toutes les autres étiquettes obtiennent un score de confiance nul.

4.2 Fusion par classification

Afin d'exploiter au mieux l'information fournie par les différents systèmes de classification développés pour ces tâches, appelés ici systèmes initiaux, nous avons mis en place un mécanisme de prise de décision basé sur des classifieurs.

4.2.1 Architecture

Nous avons opté pour une architecture originale à deux niveaux au-dessus des systèmes initiaux, représentée dans la figure 1. Le premier niveau consiste en un ensemble de classifieurs qui sont entrainés séparément à classer les documents en prenant en entrée les supervecteurs rassemblant les prédictions des systèmes initiaux. Pour chaque classe, ces classifieurs fournissent un score de prédiction tenant compte des résultats fournis par les systèmes initiaux. Le second niveau est constitué d'un classifieur qui va prendre la décision finale à partir des prédictions fournies par le premier niveau. L'entrée de ce dernier classifieur est un supervecteur constitué des scores de prédiction pour chaque classe produit par le premier niveau de classification.

Les classifieurs du premier niveau de classification peuvent être vus comme des angles d'observation différents, chacun étant adapté à seulement une partie des documents. Nous espérons ainsi que pour chaque document au moins un classifieur propose la bonne décision. Le classifieur du second niveau sert alors à prendre la décision finale en fonction des prédictions de ces points de vue complémentaires.

4.2.2 Paramètres

Notre système de fusion par classification est conçu pour recevoir en entrée un score par classe et par document pour chaque système à fusionner. Techniquement, les scores de prédiction obtenus pour chacune des classes par les différents systèmes sont concaténés dans un supervecteur de paramètres qui sera fourni à notre système de fusion par classification. Les systèmes que nous avons inclus sont tous les systèmes élémentaires décrits dans cet article, auxquels s'ajoute le système de fusion par vote majoritaire. En effet, un système de fusion reste un système et fournit des prédictions au même titre que les systèmes initiaux, mais ses scores ont la particularité de contenir une information globale sur l'ensemble des systèmes.

4.2.3 Classifieurs pour les 2 niveaux

Nous avons sélectionné 21 classifieurs de fusion de premier niveau. Parmi ceux proposés par l'outil WEKA [4], ont été retenus ceux qui donnaient les meilleurs résultats sur le corpus d'entraînement. Il y a donc 8 arbres de décision, un réseau bayesien, 3 SVM, 3 régressions, 2 votes et 4 classifieurs basés sur des règles. Les multiples instances de classifieurs au sein d'une famille correspondent à différentes implémentations ou différents réglages fournissant des résultats différents.

Le classifieur de niveau 2 retenu est celui qui fournissait les meilleures performances sur le corpus d'apprentissage : une régression logistique.

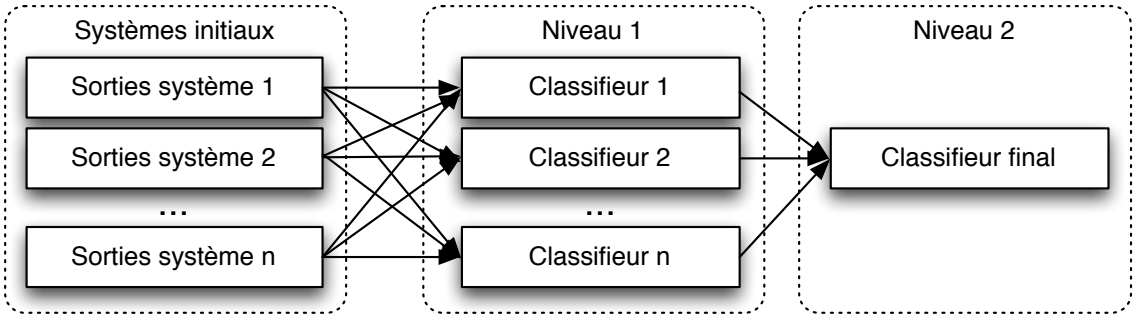


FIGURE 1 – Architecture du système de fusion par classification

5 Évaluation

5.1 Évaluation stricte

Comme décrit dans la section 2, le but du défi est de classer des documents dans un nombre fini de classes. Pour la tâche 1, il s'agit de les classer suivant leur décénnie de rédaction et pour la tâche 2 suivant le pays ou le journal de publication. Pour chaque tâche, un corpus d'apprentissage est mis à notre disposition afin de développer les algorithmes. Ceux-ci sont évalués sur des corpus de test (T) avec des caractéristiques semblables à celui d'apprentissage, en calculant la *F-mesure* des documents bien classés, moyenné sur tous les corpus :

$$F - mesure(\beta) = \frac{(\beta^2 + 1) \times \langle Précision \rangle \times \langle Rappel \rangle}{\beta^2 \times \langle Précision \rangle + \langle Rappel \rangle} \tag{3}$$

où la précision moyenne et le rappel moyen sont calculés comme :

$$\langle Précision \rangle = \frac{\sum_{i=1}^{n} Précision_i}{n} \; ; \; \langle Rappel \rangle = \frac{\sum_{i=1}^{n} Rappel_i}{n} \tag{4}$$

Étant donné pour chaque classe i :

$$Précision_i = \frac{\{\text{Nb de documents correctement attribués à la classe} i\}}{\{\text{Nb de documents total attribués à la classe} i\}} \tag{5}$$

4. http://www.cs.waikato.ac.nz/~ml/weka/

$$Rappel_i = \frac{\{\text{Nb de documents correctement attribués à la classe}i\}}{\{\text{Nb de documents appartenant à la classe}i\}} \quad (6)$$

5.2 Indice de confiance pondéré

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 la probabilité pour un document d'appartenir à une classe d'opinion donnée. La F-mesure pondérée par l'indice de confiance a été utilisée pour l'évaluation des systèmes soumis à DEFT'10 si un indice de confiance était fourni par les participants. Dans la F-mesure pondérée, la précision et le rappel pour chaque classe ont été pondérés par l'indice de confiance. Ce qui donne :

$$Précision_i = \frac{\sum_{\text{AttribuéCorrect}_{i=1}}^{\text{NbAttribuéCorrect}_i} \text{Indice\_confiance}_{\text{AttribuéCorrect}_i}}{\sum_{\text{Attribué}_{i=1}}^{\text{NbAttribué}_i} \text{Indice\_confiance}_{\text{Attribué}_i}} \quad (7)$$

$$Rappel_i = \frac{\sum_{\text{AttribuéCorrect}_{i=1}}^{\text{NbAttribuéCorrect}_i} \text{Indice\_confiance}_{\text{AttribuéCorrect}_i}}{\{\text{Nb de documents correctement attribués à la classe } i\}} \quad (8)$$

avec
– $\text{NbreAttribuéCorrect}_i$: nombre de documents appartenant effectivement à la classe i et auxquels le système a attribué un indice de confiance non nul pour cette classe.
– NbreAttribué_i : nombre de documents attribués auxquels le système a attribué un indice de confiance non nul pour la classe i.
Dans le cadre de DEFT'10, le calcul de la F-mesure retenue par les organisateurs est celui de la formule 3 avec les précision et rappel des formules 7 et 8. Cette réécriture suppose évidemment que β soit égal à 1 de façon à ne pas privilégier la précision ou le rappel.

5.3 Discussion

Pour la tâche 1, étant donné qu'il s'agit de positionner des documents sur une échelle continue de temps, le simple contrôle de l'appartenance à une décennie semble un peu brutal : un article écrit en décembre 1899 est considéré comme faux s'il est étiqueté 1900. Le calcul de l'erreur pourrait tenir compte de l'éloignement "temporel" entre la décennie donnée et celle de référence, ou bien utiliser des classes avec recouvrement et considérer comme juste les deux classes possibles en cas de chevauchement.

De plus, la méthode d'évaluation par F-mesure pondérée est utilisée systématiquement pour l'évaluation de DEFT'10 par les organisateurs. De notre point de vue, il serait plus judicieux de présenter les deux. En effet dans la mesure où nous n'avons aucun contexte précis pour utiliser les scores de confiance fournis par les systèmes seul le maximum de leur distribution peut être utilisé d'un point de vue opérationnel. Par ailleurs, n'étant pas informés de ce mode d'évaluation nous n'avons pas tenté d'optimiser nos scores de confiance pour maximiser la F-mesure pondérée comme on aurait pu le faire par exemple en appliquant une fonction de transformation des scores de confiance (*mapping*).

6 Résultats

Afin de ne pas surcharger de chiffres cet article, nous avons choisi de ne noter que la F-mesure finale de chaque système, obtenue pour chaque tâche. Plus de détails sont donnés uniquement pour le meilleur système. Les résultats sont donnés dans les tableaux 6 et 7.

	Apprentissage			Test			Test (F-m pondérée)		
Systèmes	Décennies	Journal	Pays	Décennies	Journal	Pays	Décennies	Journal	Pays
Mick_MLP	29,0	-	-	33,8	-	-	19,0	-	-
Mick_SVM	30,3	71,7	91,9	31,3	74,8	92,6	25,9	69,8	90,3
Jmt	-	67,3	87,9	-	68,3	89,6	-	68,3	89,6
Jmt_basic	-	68,4	88,9	-	68,3	90,0	-	68,2	90,0
Boost_basic	25,6	79,3	95,0	26,2	83,0	96,6	8,3	37,9	82,0
Rk_icsiboost	20,8	74,9	94,2	23,9	74,3	94,6	21,5	72,7	94,7
Hybride_boost_MLP	28,8	-	-	29,0	-	-	24,6	-	-
Fusion majoritaire	33,3	79,1	95,4	34,3	80,4	96,3	29,4	74,1	93,2
Fusion par classification	36,1	84,0	96,9	36,3	83,0	97,8	26,5	70,5	96,4

TABLE 6 – Ensemble des résultats obtenus par les systèmes initiaux et après fusion.

On remarque que tous les systèmes ont leur performance affectée par le passage au scoring avec pondération. Le système le plus touché est *Boost_basic* qui chute sur le corpus de test de 26,2% à 8,3% pour la tâche Décennie, de 83% à 38% pour les Journaux et de 96,6% à 82% pour les Pays. Ceci s'explique par le fait que les scores fournis ne représentent pas la probabilité de décision pour chaque étiquette mais un score de confiance compris entre 0 et 1 pour chaque décision indépendamment des autres étiquettes. La plupart des scores fournis sont compris entre 0,3 et 0,7. Il va de soi que ces scores auraient été paramétrés différemment si l'on avait eu conscience de leur importance dans le calcul des résultats. En revanche on remarque que ces systèmes sont robustes et obtiennent parfois de meilleurs résultats sur le corpus de test quelle que soit la tâche.

Pour le système Rk_icsiboost, les résultats obtenus sur la tâche Origine géographique sont d'excellentes qualités pour la **Sous-tâche 1**. Pour la **Sous-tâche 2**, ceux-ci restent corrects malgré un nombre conséquent d'erreurs, principalement des documents mal classés entre les deux journaux canadiens (D et P) selon la matrice de confusion. Nous attribuons la faiblesse des résultats obtenus sur la tâche Décennie au bruit produit par les erreurs d'OCR récurrentes dans le corpus. On observe cependant une bonne robustesse du système sur l'ensemble des deux tâches puisque les résultats obtenus sur les corpus de développement et celui de test sont proches.

Concernant le système *Hybride_boost_MLP*, il obtient des résultats honorables pour la tâche Decennies mais n'est pas le plus performant des systèmes initiaux et donc mériterait de meilleurs résultats vis à vis des efforts fournis (14 systèmes différents au total !).

Les modèles de n-grammes de lettres ont été testés uniquement sur la tâche Origine (sous-tâches Journal et Pays). Les performances en terme de F-mesure sur les ensembles de developpement sont autour de 89% pour les Pays et 68% pour les Journaux, le modèle "basic" étant légèrement meilleur que l'autre. Dans les deux cas, les deux types de modèles se sont avères relativement stables en deveppement et en test (90% pour les Pays et 68,3%).

La fusion majoritaire obtient toujours de meilleurs résultats que l'ensemble des systèmes initiaux mais est

surpassée par la fusion par classification. Là encore, on remarque qu'au niveau du calcul des scores pondérés la relation s'inverse et la fusion majoritaire obtient un meilleur résultat que la fusion par classification.

Concernant la fusion par classification, il est intéressant de mesurer le gain apporté par l'architecture à deux niveaux proposée. On peut le faire en comparant les performances du meilleur classifieur de niveau 1 et les performances du niveau 2. Ces résultats se trouvent respectivement dans les lignes "Meilleur niveau 1" et "Score niveau 2" du tableau 7. On constate une diminution de 0,1% absolu du taux de classification pour la tâche Pays, aucun changement pour la tâche Journal mais un gain de 1% absolu pour la tâche Décennie. On peut donc dire que cette architecture apporte un gain par rapport à une fusion utilisant un seul classifieur pour la tâche Décennie et qu'elle ne dégrade globalement pas les performances obtenues pour les autres tâches.

Afin de mesurer le gain global apporté par cette architecture, nous avons comparé les performances du meilleur système initial avec les résultats fournis par le second niveau de classification. La ligne intitulée "Meilleur système" du tableau 7 contient le taux de classification obtenu par le meilleur système initial et la ligne "Score niveau 2" contient le taux de classification final obtenu par la combinaison. On constate que la combinaison apporte une amélioration systématique par rapport au meilleur système initial sur les trois tâches considérées.

	Apprentissage			Test		
Mesure	Décennie	Journal	Pays	Décennie	Journal	Pays
Meilleur système	30,0	81,2	94,8	34,3	80,8	96,5
Oracle systèmes	64,8	94,6	99,9	64,5	94,6	99,9
Meilleur niveau 1	35,1	84,0	97,0	36,4	83,3	97,8
Oracle niveau 1	81,5	96,1	99,6	81,6	95,7	99,5
Score niveau 2	36,1	84,0	96,9	36,3	83,0	97,9

TABLE 7 – Performances en terme de F-mesure pour les trois tâches de classification considérées sur les corpus d'apprentissage et de test, mesurées aux différents niveaux du système de fusion par classification.

Afin de mesurer l'évolution des performances de la fusion par classification nous avons refait sur le corpus de test les mesures que nous avions faîtes sur le corpus de développement et qui nous avaient permis de valider l'approche. La partie droite du tableau 7 contient les résultats de ces mesures. On constate que les résultats obtenus sur les trois tâches à la sortie de la fusion (ligne "Score niveau 2") sont meilleurs que ceux du meilleur des systèmes initiaux (ligne "Meilleur système"), ce qui signifie que la fusion par classification permet toujours d'améliorer les résultats. L'architecture de classification à deux niveaux que nous avons utilisé permet donc de tirer parti de la complémentarité des classifieurs constituant le niveau 1, et qu'elle permet un gain de classification important sur la tâche Décennie par rapport à une architecture plus simple. Globalement, la fusion par classification que nous avons proposé améliore significativement les résultats obtenus par le meilleur système initial.

7 Conclusion et perspectives

La classification de documents est une tâche qui peut être très difficile en fonction du type de textes. Comme cela avait constaté lors des défis précédents, *"La lecture directe ne suffit pas toujours pour se forger un avis et privilégier une classification par rapport à une autre."* (Torres-Moreno *et al.*, 2009). Comme dans le passé, nous avons utilisé des approches de représentation numériques et probabilistes, afin de rester aussi indépendant que possible des sujets traités. Concernant les systèmes de base, Rk_icsiboost obtient de bonnes performances générales sur la tâche Origine mais se heurte aux erreurs d'OCR sur la tâche Décennie qui introduit du bruit dans le modèle en n-grammes de mot. Les modèles de n-grammes s'avérent intéressants dans la mesure où ils sont combinés avec d'autres méthodes, car ils permettent de capturer certaines caractéristiques très fines des documents. Nous pensons les améliorer en les combinant avec l'approche probabiliste de Bayes et en utilisant des mesures de distance non linéaires. Enfin nous avons présenté deux stratégies de fusion de méthodes qui se sont avérées robustes et performantes, dans tous les cas au-dessus des moyennes des meilleures soumissions initiales.

Références

BÉCHET F., EL-BÈZE M. & TORRES-MORENO J.-M. (2008). En finir avec la confusion des genres pour mieux séparer les thèmes. In *DEFT'08*, p. 161–170.

CAVNAR W. & TRENKLE J. (1994). N-gram-based text categorization. *Ann Arbor MI*, **48113**, 4001.

CHARTON, ERIC, CAMELIN, NATHALIE, ACUNA-AGOST, RODRIGO, GOTAB, PIERRE, LAVALLEY, REMI, KESSLER, REMY, FERNANDEZ & SILVIA (2008). Pré-traitements classiques ou par analyse distributionnelle :application aux méthodes de classification automatique déployées pour deft08. In *DEFT'08*, Grenoble, France.

EL-BÈZE M., TORRES-MORENO J.-M. & BÉCHET F. (2005). Peut-on rendre automatiquement à César ce qui lui appartient ? Application au jeu du Chirand-Mitterrac. In *DEFT'05*, volume 2, p. 125–134.

EL-BÈZE M., TORRES-MORENO J.-M. & BÉCHET F. (2007). Un duel probabiliste pour départager deux présidents. *RNTI E-10*, p. 117–126.

FREUND Y. & SCHAPIRE R. E. (1996). Experiments with a new boosting algorithm. In *Thirteenth International Conference on Machine Learning*, p. 148–156.

GRAVIER G., BONASTRE J. F., GEOFFROIS E., GALLIANO S., MC TAIT K. & CHOUKRI K. (2004). The ESTER evaluation campaign of rich transcription of French broadcast news. In *LREC*, p. 885–888.

MANNING C. D. & SHÜTZE H. (2000). *Foundations of Statistical Natural Language Processing*. Cambridge, Massashusetts : MIT Press.

SCHAPIRE R. E. & SINGER Y. (2000). BoosTexter : A boosting-based system for text categorization. *Machine Learning*, **39**, 135–168.

STANISLAS OGER, MICKAEL ROUVIER G. L. (2010). Transcription-based video genre classification. In *ICASSP*, p. 5114–5117.

TORRES-MORENO J. M., EL-BÈZE M., BÉCHET F. & CAMELIN N. (2007). Comment faire pour que l opinion forgée à la sortie des urnes soit la bonne ? application au défi deft 2007. In *DEFT'07*, p. 119–133.

TORRES-MORENO J. M., EL-BÈZE M., BÉCHET F. & CAMELIN N. (2009). Fusion probabiliste appliquée à la détection et classification d'opinions. In *DEFT'09*.

Décennie d'un article de journal par analyse statistique et lexicale

Pierre ALBERT Flora BADIN Maxime DELORME Nadège DEVOS
Sophie PAPAZOGLOU Jean SIMARD[1]
(1) CNRS–LIMSI, 91403 ORSAY
flora.badin@limsi.fr

Résumé. Dans le cadre du DÉfi de Fouille de Texte (DEFT) 2010, nous avons proposé une méthode permettant de dater des articles de journaux. L'étude du corpus a permis de relever les différentes caractéristiques des articles identifiables de façon automatique telles que la présence d'entités nommées et de variations orthographiques. Une approche mixte se basant à la fois sur ces propriétés linguistiques et sur les statistiques a été développée. Grâce à cette approche, les résultats ont permis d'obtenir une F-mesure allant jusqu'à 0.338.

Abstract. For the 2010 edition of DEFT, we proposed a method for dating newspaper articles. Studying the corpus allowed us to extract distinctive features of the articles. These features such as named entities and orthographic variations are automatically identificable. We developed a mixed approach based on the recognition of those linguistic properties and on statistics. Thanks to this method we have been able to achieve F-measures up to 0.338.

Mots-clés : Analyse statistique, entités nommées, fouille de texte, algorithme d'apprentissage.

Keywords: Statistical analysis, named entities, text-mining, learning algorithms.

1 Introduction

La fouille de texte consiste à extraire, en s'aidant de théories linguistiques, une information précise d'un document. Ainsi, la tâche proposée par le concours DEFT 2010 a pour objectif la datation d'articles de journaux. La connaissance de la date d'un document est très important en Traitement Automatique des Langues (TAL) puisqu'il permet de replacer le document dans son contexte historique afin de mieux l'interpréter.

L'utilisation de GALLICA, d'où sont extraits ces articles, est la seule ressource non-autorisée. Si un minimum de traitement logique est possible, notamment à travers les entités nommées, il est aussi intéressant de s'intéresser aux marqueurs caractéristiques d'une époque. Nonobstant l'efficacité de ces traitements, combiner ceux-ci avec un apprentissage statistique est indispensable pour éviter les biais et les absences d'information. Pour ce dernier, deux approches ont été utilisées, l'une portant sur la fréquence d'apparition des mots et l'autre sur leur enchainement. Par l'association de ces méthodes, nous avons estimé les décennies plausibles pour les extraits proposés. Nos résultats montrent la pertinence de cette approche, qui pourrait cependant bénéficier de développements supplémentaires afin d'être réellement efficace.

Dans une première section, nous revenons sur les caractéristiques du corpus d'entraînement qui nous a été fourni dans le cadre du concours. La seconde section expose les différentes solutions que nous avons retenues pour dater les articles. Une brève troisième section décrit la méthode de soumission de nos résultats. La dernière section permet de conclure et propose quelques évolutions possibles qui pourraient s'ajouter aux méthodes existantes ainsi que des perspectives qui pourraient être intéressantes.

2 Corpus

Durant ce DÉfi Fouille de Texte 2010, un corpus de plus de 6315 articles de journaux a été mis à notre disposition. Environ 60 % a été annoté et nous a servi de corpus d'entraînement. Le reste du corpus a été utilisé pour l'évaluation de notre modèle et a donné lieu à la soumission pour le concours DEFT 2010. Nous allons ici décrire plus en détails le corpus d'entraînement. Puis nous reviendrons sur l'ensemble des caractéristiques de ce corpus qui ont orienté notre réflexion vers une méthode d'analyse statistique et lexicale.

2.1 Le corpus d'entraînement

Le corpus d'entraînement est composé d'articles de journaux parus entre 1800 et 1944. Tous les articles sont issus de deux journaux : *La Croix* et le *Journal des Débats et des Décrets* devenu ensuite le *Journal de l'Empire* puis le *Journal des Débats Politiques et Littéraires*. Les sujets qui y sont traités sont très variés, allant de l'actualité internationale (concernant les guerres par exemple) à l'abolition de lois en passant par le simple fait divers (comme la météorologie). En dehors du journal *La Croix*, les articles sont des résumés de débats ayant eu lieu dans divers corps administratifs ou gouvernementaux. Ces articles ont été numérisés par reconnaissance optique de caractères (OCR pour *Optical Character Recognition*) à partir de la version papier des journaux puis découpés en 3594 articles. Ils ont ensuite été structurés au format XML pour les besoins de DEFT 2010. La structure XML donne pour chaque article les informations sur le nom du journal, la date de publication et la décennie. Toutes les années pouvant apparaître dans le corps

de texte des articles ont été masquées par des balises `<annee />`.

2.2 Caractéristiques du corpus

Une annotation manuelle d'une cinquantaine d'articles a permis de constater un ensemble de caractéristiques propres au corpus.

La reconnaissance optique de caractères est une technique encore imparfaite ne permettant pas de numériser un texte sans erreur. De plus, l'âge [1], la qualité du papier ou de l'encre, la présence de pliures, de taches ou même de traits de présentation (pour séparer les colonnes par exemple) sur les journaux de notre corpus ont augmenté de façon considérable le nombre d'erreurs commises par le moteur de reconnaissance optique de caractères. Celles-ci sont de natures diverses : erreur d'identification du caractère (*lecteyrs*), insertion d'espace dans un mot (*pro position*) ou omission d'espace entre deux mots (*laquestion*), trait ou pliure verticale sur le journal identifié en tant que caractère (*jour |**i** est demandé*) ou pliure horizontale perturbant la reconnaissance optique sur l'ensemble d'une ligne (*apeix:evoir*).

Ces erreurs ont mené à plusieurs pistes, certaines exploitées et expliquées dans la suite de ce papier, d'autres non-exploitées.

Par exemple, certaines erreurs se sont révélées être liées aux réformes de l'orthographe durant la période concernée par les articles (voir section 3.1).

Une autre piste aurait pu consister à essayer de trouver une corrélation entre le nombre, la fréquence ou le type d'erreurs dans un texte avec la période de l'article. Mais cela suppose de pouvoir identifier les erreurs ce qui est une tâche très difficile. Cette piste n'a donc pas été creusée.

Nous avons également constaté que les versions numériques des articles contiennent des doubles espaces à intervalles réguliers. Ces doubles espaces correspondent aux retours à la ligne en fin de colonnes. Un outil a été développé pour étudier la corrélation entre la largeur des colonnes, les noms des journaux et l'année de parution des articles. Malheureusement, ce travail n'a pas donné de résultat suffisamment pertinent pour être exploitable.

Les caractéristiques de ce corpus, combinées au faible nombre de mots par article (300 mots), ne favorisent pas l'utilisation de méthodes habituelles en Traitement Automatique des Langues (TAL). C'est pour ces raisons que nous avons délibérément évité toute étude sémantique des textes. Nous nous sommes plus concentrés sur des approches statistiques et lexicales.

3 Les méthodes

3.1 Les réformes de l'orthographe

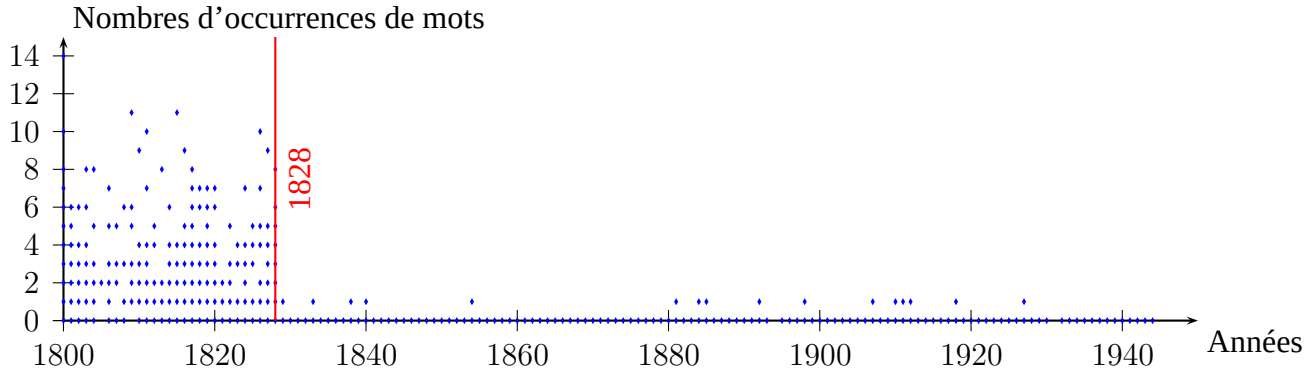
Concernant la période couverte par le corpus, deux réformes de l'orthographe ont eu lieu. Tout d'abord, celle de 1835 (voir [Académie française, 1835]) puis celle de 1878 (voir [Académie française, 1878]).

1. N'oublions pas que certains de ces articles ont plus de 200 ans.

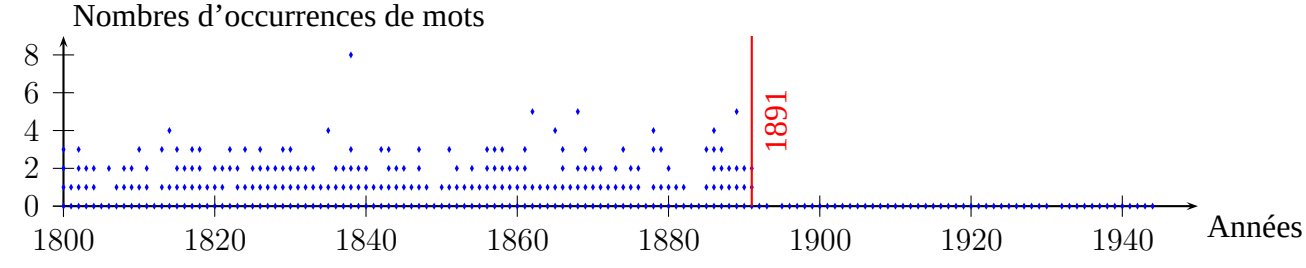
L'annotation manuelle de quelques articles nous a permis d'identifier deux éléments récurrents de la réforme de 1835 :
– La combinaison de lettres *oi* s'est vue transformée en *ai* dans une grande majorité des mots ;
– Le pluriel des mots en *nt* est passé de *ns* à *nts*.
Une étude a été réalisée sur le corpus d'entraînement afin de vérifier que ces deux règles fournissent une bonne indication concernant la date de l'article en question. Pour cela, un algorithme simple mais relativement efficace a été appliqué :

1. Filtrer tous les mots se terminant par *ois*, *oit* et *oient* ;
2. Dans cette liste de mots, retirer tous les mots se trouvant dans le dictionnaire contemporain de la langue française (comme *trois*, *aperçois*, *voient*) ;
3. Parmi les mots restants, remplacer la lettre *o* dans *ois*, *oit* et *oient* par la lettre *a* ;
4. Vérifier que le nouveau mot se trouve dans le dictionnaire contemporain de la langue française :
   – Si le mot se trouve dans le dictionnaire, il est comptabilisé ;
   – Si le mot ne se trouve pas dans le dictionnaire, il est enlevé de la liste.

Concernant les pluriels des mots se terminant par *nt*, le même algorithme a été utilisé. L'application de ces deux algorithmes sur le corpus d'entraînement a permis d'obtenir un couple (A, N) pour chaque article où A est l'année de l'article et N est le nombre d'occurrences constatées concernant la réforme dans ce même article. L'ensemble des couples répertoriés sont représentés sur la Figure 1 (un point sur le graphique pouvant représenter plusieurs articles). Ils permettent de mettre en évidence deux limites distinctes.



(a) La terminaison des mots en *oi* (réforme de 1835)



(b) Le pluriel des mots terminant par *nt* (réforme de 1878)

Figure 1 – Les réformes de l'orthographe.

La mesure de ces deux indices révèle avec certitude deux filtres sur les articles.

Le premier filtre concerne les mots dont la terminaison contient *oi*. Pour chaque article contenant plus de deux occurrences, la probabilité que la date de l'article soit supérieure à 1828 est nulle. Dans le cas où l'article contient une occurrence, la probabilité que la date de l'article soit supérieure à 1828 est $17/187$ soit 0.091 (proportion d'articles contenant une occurrence et dont la date est supérieure à 1828 par rapport au total des articles contenant une occurrence).

Le second filtre concerne les mots au pluriel et se terminant par *ns*. Pour chaque article contenant au moins une occurrence, la probabilité que la date de l'article soit supérieure à 1891 est nulle.

Dans le cas où aucun des deux indices n'est détecté dans l'article, aucun filtre n'est appliqué.

3.2 Les entités nommées

Concernant les entités nommées, les articles sont en général assez fournis. Nous trouvons aussi bien des entités nommées du type *M. Dupond*, *Napoléon III* que *France* ou *boulevard Haussmann*. Étant donnée la casse particulière des entités nommées, ce sont potentiellement des groupes de mots facilement identifiables dans un article.

Ce sont les noms propres désignant une personne qui ont été choisis comme axe de recherche. En effet, ils contiennent plusieurs caractéristiques (voir [Ehrmann, 2008]) :
– Il est possible d'associer une date de naissance à un nom de personne ce qui permettrait de filtrer les années possibles de l'article ;
– Le nom d'une personnalité est souvent accompagné de préfixes distinctifs tels que *M.*, *docteur* ou de suffixes tel qu'un chiffre romain (*Napoléon III*).
Voici une liste non-exhaustive des préfixes permettant d'identifier un nom : *M.*, *Madame*, *Mlle*, *Président*, *EVEQUE*, *Lord*... Concernant le nom propre, des préfixes ont également été identifiés tels que *Van den* comme dans *Van den Berghe* ou *de la* comme dans *de la Fontaine*. Pour les suffixes, seuls les chiffres romains ont été identifiés.

Si un ou plusieurs noms sont relevés dans un article, chaque nom fait l'objet d'une recherche sur Internet par le biais du site UNIVERSALIS afin de trouver une date de naissance. Bien évidemment, un résultat n'est pas systématiquement trouvé. Parfois, le résultat peut également être inutile. Par exemple, si le nom *Aristote* est trouvé dans un article, il est évident que sa date de naissance ne nous permettra pas d'en déduire une année pour un article datant de la période 1800–1944.

Si une date de naissance est trouvée, une faible probabilité sera donnée aux années antérieures à cette date. Dans notre cas, cette probabilité est 0.2. Ensuite, nous avons estimé que la probabilité qu'une personne soit citée dans la presse avant l'âge de ces 20 ans est relativement faible mais qu'elle augmentait linéairement avec l'âge. Nous avons donc une probabilité linéairement montante entre la date de naissance de la personne et la date de ces 20 ans. Après ces 20 ans, c'est une probabilité de 0.8 qui est donnée (voir Figure 2 page suivante).

3.3 Apprentissage par l'utilisation de *Conditional Random Fields*

Afin de pallier au manque de précision des outils lexicaux, une étude statistique par apprentissage, fondée en partie sur la méthode récente des *Conditional Random Fields* (CRF), est utilisée [Lafferty *et al.*, 2001].

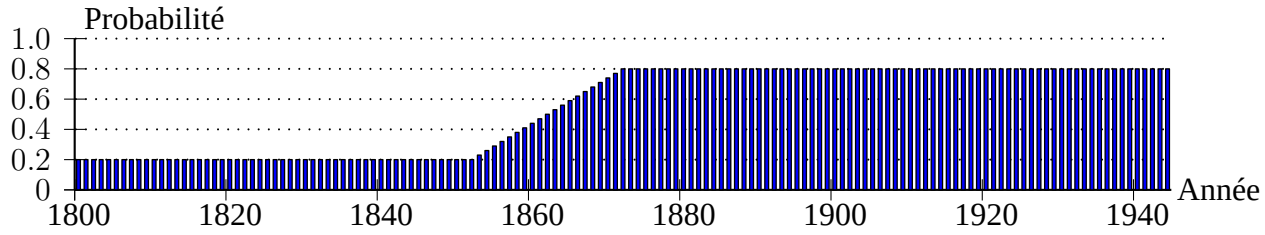


Figure 2 – Probabilité appliquée sur un article en fonction d'une date de naissance : Exemple avec Calamity JANE (1852–1903).

Dans ce cadre, un article est perçu comme un graphe linéaire de mots. Chaque mot est un nœud relié à une étiquette qui le caractérise (dans ce cas, sa date). L'apprentissage se fait alors en calculant les probabilités de transition du mot à chaque étiquette en fonction de son contexte. Le principe des CRF est voisin de celui des automates de MARKOV à états cachés. Dans ce dernier, la séquence de l'automate n'est pas connue et seul le résultat est visible. Le CRF en est une généralisation non-orientée, avec un nombre de contraintes illimitées, et dont les probabilités de transition varient en fonction de la séquence analysée [Wallach, 2004]. Les CRF sont particulièrement intéressants pour le traitement automatique des langues. Dans le cadre de l'étude de la variation diachronique, l'étiquetage se fait sur la période (la décennie dans notre cas). Parmi les différentes implémentations de l'algorithme, CRF++ est adapté au cadre des graphes linéaires.

La détermination de la date d'un mot n'étant pas liée aux mots qui l'entourent mais à leurs étiquettes, la règle d'entraînement est constituée des cinq unigrammes simples suivants :
– $U\%x[-2,0]$;
– $U\%x[-1,0]$;
– $U\%x[0,0]$;
– $U\%x[1,0]$;
– $U\%x[2,0]$;
Nous observons le contexte d'un mot sur une fenêtre totale de cinq mots et de leurs étiquettes.

3.3.1 Traitement du corpus

Le corpus a été découpé afin de prendre en compte les mots ainsi que la ponctuation qui a été ajoutée dans un second temps suite aux faibles résultats d'une première évaluation. L'apport de données sur ce corpus de taille relativement faible a eu une incidence non-négligeable. Les mots reconnus comme mal numérisés ont aussi été conservés afin de prendre en compte les erreurs récurrentes et potentiellement des schémas caractéristiques d'une période, même si ce dernier point semble avoir donné peu de résultats au regard de nos observations. L'apprentissage se faisant sur les mots et non sur leurs lettres, les erreurs ponctuelles n'ont aucune incidence. Il serait intéressant d'observer le gain apporté par une correction des erreurs de numérisation, comme nouvel apport d'informations justes.

L'étiquetage retenu est celui des décennies, limitant le nombre d'étiquettes possibles à quinze. Potentiellement plus fin, il a été favorisé par rapport à celui des années qui entraîne une multiplication des étiquettes (145 possibilités). En plus de diminuer les coûts d'apprentissage, cela permet de disposer virtuellement d'un plus grand corpus pour chaque étiquette. Le gain en précision qu'aurait apporté une première passe en année (regroupement de probabilités) est ici largement contrebalancé.

3.3.2 Expérimentations et résultats

L'entraînement étant particulièrement coûteux en ressources, seule une dizaine de configurations ont été testées (voir Source 1).

Source 1 – Résultats d'un apprentissage avec une fenêtre de cinq mots sur un article brut

```
resultats (total : 942):
difference  : nombre    pourcentage   cumul
ok  : 101   10.7218683651805    10.7218683651805
10  : 98    10.4033970276008    21.1252653927813
20  : 124   13.1634819532909    34.2887473460722
30  : 92    9.76645435244161    44.0552016985138
40  : 85    9.02335456475584    53.0785562632696
50  : 81    8.59872611464968    61.6772823779193
60  : 57    6.05095541401274    67.728237791932
70  : 61    6.4755838641189     74.2038216560509
80  : 53    5.62632696390658    79.8301486199575
90  : 54    5.73248407643312    85.5626326963906
100 : 46    4.88322717622081    90.4458598726115
110 : 43    4.56475583864119    95.0106157112526
120 : 20    2.12314225053079    97.1337579617834
130 : 16    1.69851380042463    98.8322717622081
140 : 11    1.16772823779193    100
```

Comparé à un système purement aléatoire, un écart significatif est relevé. L'algorithme se trouve bien au-dessus des résultats attendus pour les décennies proches. La moitié du corpus est identifié avec moins de quatre décennies d'écart (contre 37 % attendus). L'utilisation de la ponctuation a permis d'améliorer ces résultats de près de 20 %, avec cependant un temps d'entraînement quasiment doublé. Le résultat pour chaque article est retourné sous forme d'une densité de probabilité sur les quinze décennies.

3.4 Récurrence des termes

En s'inspirant de la méthode précédente, il est possible d'entraîner le système pour qu'il attribue un score différent aux articles en entrée. Le corpus d'entraînement est *déplié* de façon à ce que chaque chaîne de caractères de chaque article [2] se voit attribuer la décennie de l'article dont elle est extraite (voir Source 2).

Source 2 – Annotation automatique des chaînes de caractères du corpus d'entraînement

```
etant           1830
la              1830
representation  1830
```

2. Chaînes de caractères consécutifs séparés par au moins un espace de chaque côté.

```
la              1830
plus            1830
pure            1830
la              1830
plus            1830
noble           1830
de              1830
la              1830
pensee          1830
democratique    1830
,               1830
devrait         1830
au              1830
moins           1830
commencer       1830
```

La phase d'entraînement va attribuer à chaque chaîne de caractères trouvée dans le corpus un histogramme indiquant les occurrences de la chaîne pour chaque décennie. Le système va donc produire, à l'issue de l'entraînement, une table pour chaque chaîne de caractères rencontrée pendant l'entraînement. Chaque table contient quinze entiers (un par décennie) représentant le nombre d'occurrences rencontrées de cette chaîne pour chaque décennie. Une fois l'intégralité du corpus d'entraînement parcouru, les tables sont normalisées de façon à représenter une densité de probabilité.

Lors de la phase d'évaluation, un tableau de quinze scalaires initialisés à zéro est créé. La présence de chaque chaîne de caractères de l'article en entrée est vérifiée dans la table d'entraînement. Si la chaîne est introuvable, alors elle est ignorée. Sinon, les valeurs de la table de la chaîne sont additionnées à la table de l'article. Ainsi, une image approximative de la probabilité de répartition des chaînes de caractères de l'article est créée au fil des décennies.

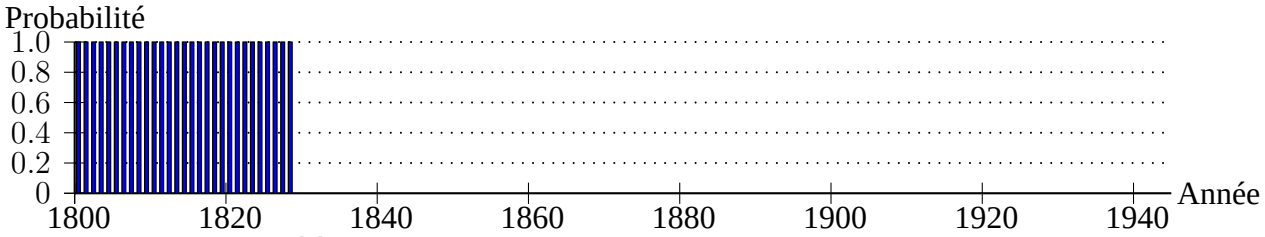
Une fois que l'intégralité d'un article est parcouru de cette manière, le tableau final est normalisé. Cette fois-ci, chaque valeur est normalisée en divisant chaque valeur du tableau par le maximum pour que chaque colonne se retrouve sur l'intervalle $[0; 1]$. Il ne s'agit plus ici d'une densité de probabilité. Ce tableau est ensuite transmis au processus de fusion qui se chargera de l'intégrer aux autres résultats.

3.5 Fusion des résultats

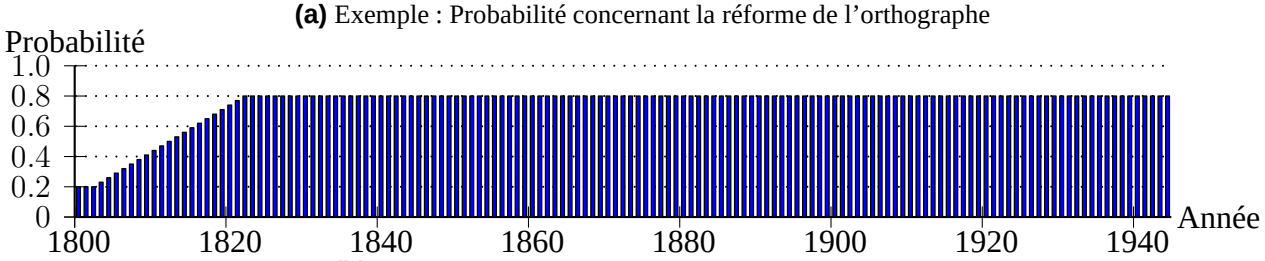
Tous les modules de traitement lexical présentés précédemment sont des filtres indépendants les uns des autres. Tous fournissent des résultats qui donnent pour un article, une probabilité sur chacune des années de 1800 à 1944. La fusion des résultats se fait par une simple multiplication des résultats de chaque filtre.

Imaginons un article dans lequel se trouveraient les mots *prouvoit* et *avoient* (respectivement *prouvait* et *avaient* en français contemporain) ainsi que la seule entité nommée *Victor Hugo*. Concernant la réforme de l'orthographe, l'apparition de deux occurrence nous permet de filtrer par rapport à l'année 1828 (voir Figure 3a page suivante). Pour les entités nommées, seule l'entité *Victor Hugo* a été identifiée et une recherche sur Internet nous permet de déterminer une date de naissance en 1802 (voir Figure 3b page ci-contre). En multipliant les deux résultats, nous obtenons la Figure 3c page suivante qui présente une

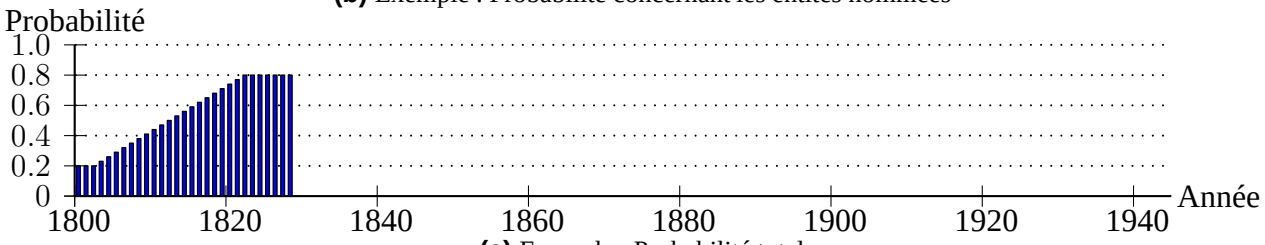
forte probabilité pour la décennie 1820.



(a) Exemple : Probabilité concernant la réforme de l'orthographe



(b) Exemple : Probabilité concernant les entités nommées



(c) Exemple : Probabilité total

Figure 3 – Exemple de fusion des résultats.

4 Soumission

La soumission finale concerne l'évaluation du corpus de test contenant 2721 textes pour lesquels aucune indication supplémentaire n'est donnée. De plus, comme pour le corpus d'entraînement, toutes les dates dans les corps de textes ont été masquées par des balises vides.

Avec notre algorithme, nous avons délibérément décidé d'appliquer certaines de nos méthodes année par année. Pourtant, le résultat final doit être une décennie ou, le cas échéant, fournir une probabilité pour chaque décennie. Le résultat doit se présenter sous la forme d'une densité de probabilité sur l'ensemble des décennies.

Nos résultats ne sont pas sous cette forme puisque une probabilité est fournie pour chaque année. De plus, la probabilité pour chaque année se trouve incluse dans l'intervalle $[0; 1]$ ce qui n'assure aucunement d'obtenir une densité de probabilité (*i.e.* la somme des probabilités n'est pas forcément égale à 1).

Pour commencer, nous avons effectué une moyenne de nos résultats par décennie.

$$p'_{\mathrm{d}} = \sum_{i=\mathrm{d}}^{\mathrm{d}+10} \frac{p_{\mathrm{d}+i}}{10} \tag{1}$$

d représentant la décennie concernée.

Puis, afin d'obtenir une vrai densité de probabilité, les probabilité de toutes les décennies ont été divisées par la somme totale des probabilités.

$$p''_{\mathrm{d}} = \frac{p_{\mathrm{d}}}{\sum_{i=1800}^{1940} p_{\mathrm{d}}} \tag{2}$$

Nous avons testé puis soumis trois différentes types de résultats puisque cela était permis dans les conditions du concours DEFT 2010. Tout d'abord, nous avons évalué la fiabilité de nos résultats avec un fichier contenant une probabilité pour l'ensemble des décennies. Ce fichier peut contenir des probabilités nulles comme dans le cas des filtres de la réforme de l'orthographe (voir section 3.1 page 3). Néanmoins, dans la plupart des cas, une probabilité non-nulle sera donnée pour chaque décennie. Ce type de résultat nous permet d'obtenir une précision de 0.297 et un rappel de 0.299. La F-mesure est de 0.298 (voir Figure 4 page suivante).

Puis nous avons tenté de voir si nous pouvions obtenir de meilleurs résultats en exploitant au mieux les sorties de notre algorithme. Par exemple, nous avons de nouveau évalué le corpus en ne conservant que la décennie ayant obtenu la plus grande probabilité. Dans le cas où plusieurs décennies sont concernées (plusieurs égalités), une probabilité égale est affectée à chacune. Cette seconde soumission donne de bien meilleurs résultats. Cette nouvelle stratégie nous a permis d'obtenir une précision de 0.336 et un rappel de 0.340 et donc une F-mesure de 0.338 (voir Figure 4 page ci-contre).

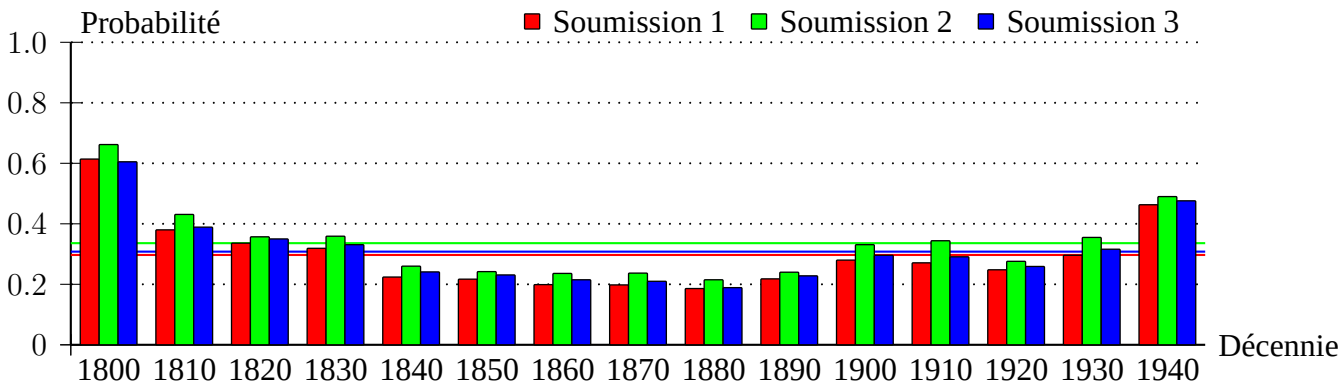
Cependant, la solution conservant uniquement la meilleure probabilité peut faire disparaître la décennie recherchée. En effet, nos méthodes étant en partie basées sur des statistiques, il est possible que la bonne décennie ne soit que la seconde voire la troisième meilleure probabilité. Nous avons donc décidé d'effectuer une troisième soumission en ne conservant que les trois meilleures probabilités. Dans le cas où plusieurs décennies possèdent la meilleure troisième probabilité, elles sont toutes conservées. La densité totale de probabilité est alors répartie sur l'ensemble des décennies concernées. Cette troisième et dernière proposition nous a permis d'obtenir une précision de 0.308 et un rappel de 0.313 et donc une F-mesure de 0.310 (voir Figure 4 page suivante).

La dernière solution est moins efficace que la seconde. Ceci est dû à la distribution des probabilités sur trois décennies (ou plus le cas échéant). Dans le cas où la seconde soumission supprimait la bonne décennie, cette troisième soumission va améliorer les résultats puisqu'elle aura plus de chance de conserver la bonne décennie. Dans le cas où la seconde soumission donnait déjà la bonne décennie, cette troisième soumission va avoir pour effet de répartir les probabilités sur deux autres décennies (ou plus) qui seront incorrectes ce qui diminue la fiabilité totale.

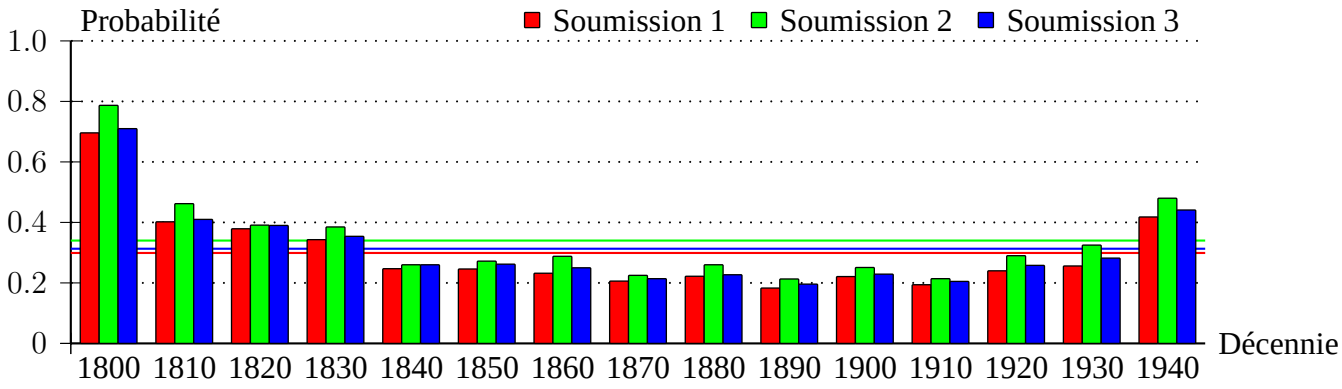
5 Conclusion

Dans cet article, nous avons présenté un système qui permet de dater des articles de journaux. Ce système se base sur différentes méthodes : méthodes basées sur le lexique, méthode statistique, apprentissage… Ces statistiques nous ont permis d'atteindre une précision de 33.6 % pour un rappel de 34 %. La F-mesure de nos résultats monte à 33.8 %.

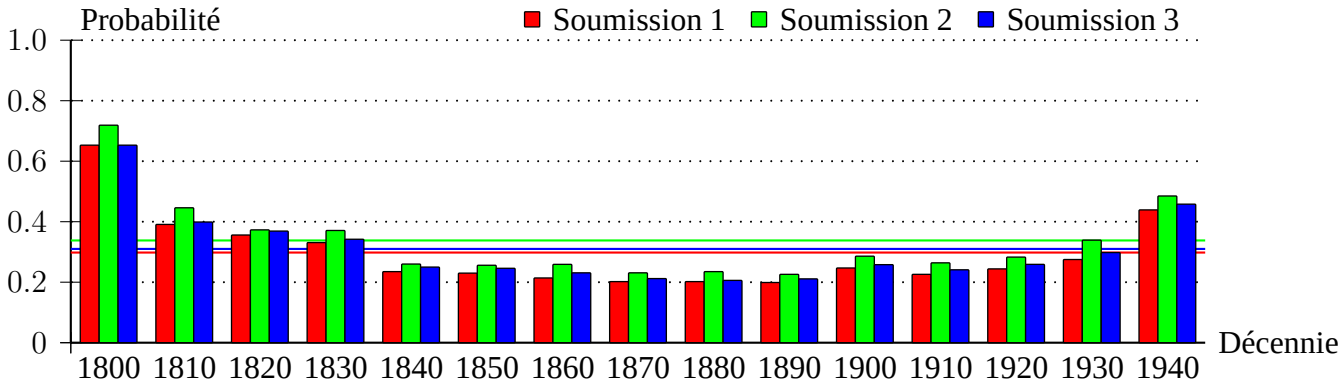
En ajoutant des modules complémentaires, nous pensons que notre système pourrait être plus efficace.



(a) Précision



(b) Rappel



(c) F-mesure

Figure 4 – Résultats des trois soumissions (précision, rappel et F-mesure).

Parmi ces modules complémentaires, certains n'ont pas pu être intégrés pour des raisons techniques alors que d'autres ont seulement été envisagés.

Afin de renforcer l'amplitude des pistes lexicales, il est logique de s'intéresser à l'étymologie de l'ensemble des mots des articles. Cette période étant très riche en inventions et technologies, le lexique associé à celles-ci a une plus grande fréquence d'apparition, principalement lors des guerres. La ressource de référence utilisée dans ce cas est le Trésor de la Langue Française Informatisé (TLFI) consulté par l'intermédiaire du site du Centre National de Ressources Textuelles et Lexicales (CNRTL). Les mots mal identifiés, présentant des caractères non-alphabétiques, sont filtrés, de même que les mots outils. Ce filtrage permet de diminuer le nombre de requêtes et ainsi diminue le temps de traitement du corpus d'environ 35 %.

Aucun traitement n'étant effectué d'un point de vue sémantique, l'ensemble des définitions sont prises en compte et la date répertoriée la plus ancienne de l'utilisation du terme est conservée. Pour un article, la date du mot le plus récent détermine ensuite avec une très grande probabilité la limite basse. Cette piste a été abandonnée en raison de problèmes techniques. Les requêtes au serveur nécessitant une charge importante, le site du CNRTL bloque notre adresse IP. Un accès local au dictionnaire ainsi qu'un cache des termes déjà traités permettrait de diminuer de façon importante le temps de recherche.

Pour améliorer le module de recherche des entités nommées, il faudrait utiliser une plus grande variété de préfixes tels que *Mgr* pour *Monseigneur* ou *S. M.* pour *Sa Majesté*. Les préfixes désignant la fonction des personnes (*Colonel* ou *Baron* par exemple) pourraient également être utilisées pour mieux filtrer les recherches sur Internet et ainsi permettre la distinction entre plusieurs personnes ayant le même nom.

Enfin, sur le modèle de la méthode des entités nommées, nous envisagerions de créer un module utilisant les dates des inventions (*macadam* ou *dynamomètre* par exemple) pour établir une limite basse. Ce module se baserait sur la supposition qu'un nom désignant une invention n'est pas utilisé tant que l'invention n'a pas été créée. Il est tout de même nécessaire de faire attention car ce module pourrait bien amener quelques erreurs comme par exemple avec le mot *voiture* qui, à l'époque, ne désignait pas une automobile mais simplement un engin tracté par des chevaux.

Toutes ces améliorations pourraient permettre d'obtenir, pour cette tâche, une meilleure F-mesure.

Références

ACADÉMIE FRANÇAISE (1835). *Dictionnaire de l'Académie française*. P. Dupont, 6e edition.

ACADÉMIE FRANÇAISE (1878). *Dictionnaire de l'Académie française*. Firmin–Didot, 7e edition.

EHRMANN M. (2008). *Les entités nommées de la linguistique au* TAL *: statut théorique et méthodes de désambiguïsation*. PhD thesis, Université PARIS VII.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional Random Fields : probabilistic models for segmenting and labeling sequence data. In [Society, 2001], p. 282–289.

I. M. L. SOCIETY, Ed. (2001). *International Conference on Machine Learning*, Princeton.

WALLACH H. M. (2004). *Conditional Random Fields : an introduction*. Rapport interne, Université de Pennsylvanie.

Index des auteurs

Albert, Pierre 85

Badin, Flora 85

Camelin, Nathalie 69

Da Sylva, Lyne 3
Delorme, Maxime 85
Devos, Nadège 85

El Ghali, Adil 51

Forest, Dominic 3

Grouin, Cyril 3
Généreux, Michel 57

Hoareau, Yann Vigile 51

Kessler, Rémy 69

Lefèvre, Fabrice 69

Mokhov, Serguei 35
Monceaux, Laura 21

Oger, Stanislas 69

Papazoglou, Sophie 85
Paroubek, Patrick 3

Rouvier, Mickael 69

Simard, Jean 85

Tartier, Annie 21
Torres-Moreno, Juan-Manuel 69

Zweigenbaum, Pierre 3

Index des mots-clés

DEFT2010 35

MARF 35

algorithme d'apprentissage 85
alida 51
analyse diachronique 57
analyse statistique 85
apprentissage automatique 69

campagne d'évaluation 3
catégorisation de documents 51
classification automatique 3
classification de textes 57, 69
comparaison des algorithmes pour le TAL 35
corpus comparables 57
corpus d'apprentissage 21
corpus d'évaluation 21
correction orthographique 57

diachronie 3
diatopie 3
défi DEFT 69

entités nommées 85
extraction d'entités 21

fouille de texte 85
frameworks 35

internationalisation 3

méthodes probabilistes 69

random indexing 51

saillance 57

variation linguistique 3