Présentation et résultats du défi fouille de texte DEFT2011 Quand un article de presse a t-il été écrit ? À quel article scientifique correspond ce résumé ?

Cyril Grouin[1] Dominic Forest[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, patrick.paroubek, pierre.zweigenbaum}@limsi.fr, dominic.forest@umontreal.ca

Résumé. Dans cet article, nous présentons l'édition 2011 du défi fouille de texte (DEFT). Pour cette édition, nous avons proposé deux tâches, l'une portant sur la variation diachronique (faisant suite à la tâche diachronique instituée lors de DEFT2010), la seconde ayant trait aux appariements entre résumés et articles scientifiques. Nous exposons dans un premier temps les tâches proposées, les modalités de constitution des différents corpus ainsi que les tests humains réalisés. Dans un second temps, nous détaillons les résultats obtenus par chacun des participants aux deux tâches. Enfin, nous concluons sur l'édition et abordons quelques pistes pour la prochaine édition du défi.

Abstract. In this paper, we present the DEFT 2011 edition. In this edition, we proposed two tasks, the first one dealing with diachronic variation (being the continuation of the diachronic variation task in DEFT2010), the second one being dedicated to the abstracts/scientific articles pairing. We first describe the proposed tasks, how corpora were created, and human evaluation. We then detail the results obtained by each participant. Finally, we made a conclusion of this edition and propose some ideas for the next challenge.

Mots-clés : Campagne d'évaluation, fouille de texte, diachronie, appariements résumés/articles.

Keywords: Evaluation campaign, Text-mining, Diachronic variation, abstracts/articles pairing.

1 Introduction

Depuis 2005, le défi fouille de texte (DEFT) propose un challenge de recherche en fouille de texte autour de thématiques régulièrement renouvellées. Depuis sa création, l'objectif du défi vise à confronter les méthodes élaborées par plusieurs équipes, sur un même jeu de données, à la manière des campagnes d'évaluation internationales qui existent dans le domaine de la recherche d'information (MUC, TREC, CLEF, etc.).

Pour cette nouvelle édition, deux appels ont été diffusés sur les principales listes de discussion dans le domaine du traitement automatique des langues *(Corpora, Humanist, LN, Risc, etc.)*, les 6 janvier et 16 février 2011. Quatorze équipes se sont inscrites, douze ayant travaillé jusqu'à la phase de test. Parmi ces différentes équipes, nous notons avec satisfaction la participation de trois équipes non francophones (FBK en Italie, INAOE au Mexique et UPF en Espagne) et à des équipes nord-américaines (Canada et Mexique). Comme en 2010, deux équipes ayant la même affiliation que les organisateurs (EBSI et LIMSI) se sont inscrites au défi. Nous précisons qu'elles n'ont bénéficié d'aucun traitement de faveur.

- CHArt, *Cognition Humaine et Artificielle*, Paris, France : Yann-Vigile Hoareau, Murat Ahat, Saïd Fouchal, Coralie Peterman et David Medernach.
- EBSI, *Ecole de Bibliothéconomie et des Sciences de l'Information*, Montréal, Canada : Romaric Boley.
- FBK, *Fondazione Bruno Kessler*, Trento, Italie : Sara Tonelli et Emanuele Pianta.
- GREYC, *Groupe de Recherche en Informatique, Image, Automatique et Instrumentalisation de Caen*, Caen, France : Gaël Lejeune, Romain Brixtel et Emmanuel Giguet.
- INAOE, *Instituto Nacional de Astrofísica, Óptica y Electrónica*, Mexico, Mexique : Fernando Sánchez-Vega, Esaú Villatoro-Tello, Antonio Juárez-González, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez et Luis Meneses-Lerín.
- IRISA, *Institut de Recherche en Informatique et Systèmes Aléatoires*, Rennes, France : Christian Raymond et Vincent Claveau.
- LIMSI, *Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur*, Orsay, France : Anne-García-Fernandez, Anne-Laure Ligozat, Marco Dinarelli et Delphine Bernhard.
- LORIA, *Laboratoire Lorrain de Recherche en Informatique et ses Applications*, Nancy, France — LASELDI, *Laboratoire de Sémiotique, Linguistique, Didactique et Informatique*, Besançon, France — Diatopie, Paris, France : Martine Cadot, Sylvain Aubin et Alain Lelu.
- LUTIN, *Laboratoire des Usages en Technologies d'Information Numérique*, Paris, France. Deux membres de ce laboratoire ont participé individuellement à DEFT2011. Nous notons ces deux participations dans la suite de ce papier LUTIN-a (Adil El Ghali) et LUTIN-d (Daniel Devatman Hromada).
- UCL, *Université Catholique de Louvain*, Louvain-la-Neuve, Belgique : Yves Bestgen.
- UPF, *Universitat Pompeu Fabra*, Barcelone, Espagne : Horacio Saggion.

Nous avons proposé aux participants de travailler sur deux tâches distinctes. La première concerne la variation diachronique en corpus de presse et attend des participants qu'ils identifient l'année de publication d'extraits de 500 mots (piste 1) ou de 300 mots (piste 2) d'articles de presse, parus entre 1801 et 1944. La seconde tâche traite du résumé d'articles scientifiques et s'articule autour de l'appariement de résumés et d'articles scientifiques complets (piste 1) et de l'appariement de résumés et du texte des articles scientifiques (piste 2, les articles ont été amputés de leur introduction et conclusion). Ces deux tâches ont porté sur des textes rédigés en français.

2 Tâche 1. Diachronie

Pour faire suite à la tâche diachronique de l'édition 2010 du défi DEFT (Grouin *et al.*, 2010) pour laquelle des méthodes originales ont été élaborées, nécessitant cependant des améliorations, nous avons décidé de proposer de nouveau une tâche sur la variation diachronique. L'édition 2010 concernait l'identification de la décennie de publication d'un extrait de 300 mots. L'un des points problématiques de ce type de tâche concerne les frontières de classe. En effet, un document paru en 1919 sera rattaché à la décennie 1910 (une décennie couvrant les années de 0 à 9 selon notre définition). Ainsi, un système retournant la décennie 1920 sera pénalisé par cette réponse, alors que l'année de parution se situe à proximité immédiate de cette décennie. Pour pallier ce problème, nous avons décidé de nous focaliser non plus sur les décennies mais sur l'année exacte de parution d'un document.

2.1 Corpus

2.1.1 Présentation générale

Le corpus a été créé à partir des archives de journaux numérisées avec reconnaissance optique de caractères mises à la disposition du public par la BNF via son portail Gallica[1]. L'ensemble des journaux ainsi numérisés a été téléchargé. Chaque fichier dans sa version texte a été découpé en portions de 500 mots et seules les portions vides de tout caractère exotique (le tilde, l'esperluette, l'accent circonflexe sans voyelle, etc.) ont été conservées. Contrairement à DEFT2010, nous n'avons travaillé que sur les pages 1 et 2 de ces journaux, partant du principe que ces pages contenaient davantage d'articles de fond que les pages 3 et 4, plutôt dévolues aux programmes du théâtre, à la bourse ou à des réclames. Nous avons conservé les proportions de l'année dernière en terme de volume pour l'apprentissage et le test. Ainsi, pour DEFT2010, 252 documents/décennie (apprentissage) et 169 documents/décennie (test) ; pour DEFT2011 : 25 documents/année (apprentissage) et 17 documents/année (test).

Deux pistes ont été proposées, l'une sur des extraits de 300 mots (comme lors de DEFT2010 afin de mesurer l'évolution des systèmes ayant participé aux deux éditions), l'autre sur des extraits de 500 mots (nouveauté de DEFT2011, de manière à voir si des améliorations sont notables avec un passage à l'échelle). Les mêmes documents ont été utilisés pour les deux pistes, les documents de 300 mots étant une partie de ceux de 500 mots (extraction aléatoire du début, du milieu ou de la fin), ordonnés différemment dans les corpus des deux pistes. Alors que l'édition 2010 du défi intégrait des extraits provenant de cinq journaux, le portail Gallica s'est entre temps enrichi de deux nouveaux titres disponibles au format texte : *La Presse* et *Le Temps*. Nous avons intégré ces deux nouveaux titres aux corpus de cette année, portant le nombre total de journaux traités à sept.

Les caractéristiques principales de ces corpus et des traitements appliqués sont les suivants :
– Les documents de travail sont le résultat d'une reconnaissance optique de caractères ; ils contiennent donc du bruit lié à cette reconnaissance de caractères (« efTorcée », « rcatisô », « cotte », « ?uf ») ;
– Nous avons éliminé les portions de journaux comportant des caractères inexistants à l'état brut en français (le tilde ˜, l'esperluette &, et le circonflexe isolé ˆ) ;
– Les caractères chevrons < et > sont remplacés par l'entité HTML correspondante (&lt ; et &gt ;) ;
– Le résultat de la reconnaissance optique de caractères ne comprenant aucun élément de structuration, les documents proposés intègrent donc des extraits d'articles incomplets (début et/ou fin manquants) ;
– Les années, lorsqu'elles étaient explicitement présentes et sans erreur dans le texte (« 1813 » par exemple, mais pas « 18 !3 »), ont été remplacées par une balise typante <annee/>.

2.1.2 Corpus d'apprentissage

Le corpus d'apprentissage intègre des articles provenant de 6 journaux différents (*Le Journal des Débats, Le Journal de l'Empire, Le Journal des Débats politiques et littéraires, Le Figaro, Le Temps* et *La Croix*), à raison d'un maximum de 25 documents par année (ce qui correspond aux 249 articles par décennie de DEFT2010). Seule l'année 1815 compte moins de 25 articles, cette année étant charnière entre deux journaux, avec trop peu de documents de qualité selon les critères fixés. La répartition des articles en fonction du journal d'origine est ici donnée (voir figure 1), sachant qu'un journal s'est prolongé dans le temps sous trois noms différents : *Le Journal des Débats* devenu *Le Journal de l'Empire* en 1805 et *Le Journal des Débats politiques et littéraires* en 1814.

2.1.3 Corpus de test

Afin d'éprouver la robustesse des systèmes, le corpus de test intègre des extraits d'articles provenant d'un septième journal absent du corpus d'apprentissage, *La Presse*. Nous avons retenu ce journal comme matériau inconnu pour deux raisons : parce qu'il n'a pas servi dans le corpus de DEFT2010 d'une part[2], et parce que sa parution a commencé plus tôt que pour *Le Temps* – autre journal absent du corpus de DEFT2010 – en 1836 au lieu de 1861. Le corpus de test propose 17 documents par année, sauf pour 1815 pour les mêmes raisons que celles évoquées dans la présentation de la constitution du corpus d'apprentissage.

[1] `http://gallica.bnf.fr/`

[2] En 2010, nous avions utilisé des articles provenant des quotidiens *Le Journal des Débats, Le Journal de l'Empire, Le Journal des Débats politiques et littéraires* et de *La Croix* pour le corpus d'apprentissage ; le corpus de test comprenait en plus des articles du quotidien *Le Figaro*.

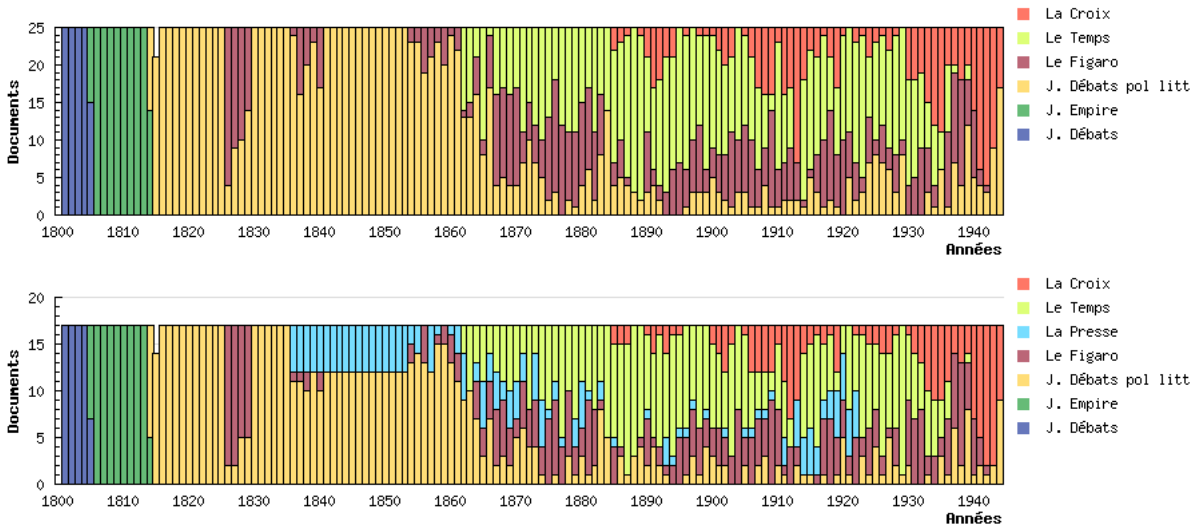

FIG. 1 – Nombre d'articles par journal et par année dans les corpus d'apprentissage et de test

2.2 Méthodes d'évaluation

Puisque l'identification d'une année au lieu d'une décennie multiplie par dix le nombre de classes à traiter d'une part, et pour ne pas faire chuter les résultats d'autre part, nous avons décidé d'évaluer les résultats sur la base d'une fenêtre de 15 ans autour de l'année de référence. Étant donné un fragment d'article a_i dont la date de parution indiquée dans la référence est $d_r(a_i)$, un système prédit une date de parution $d_p(a_i)$. Le système reçoit pour cette tâche un gain qui est d'autant plus grand que l'année prédite est proche de l'année de référence.

– Nous définissons ce gain comme une similarité entre date prédite et date de référence. Il varie entre 0 (pire) et 1 (meilleur). Notation : $s(d_p, d_r)$. Le principe d'une similarité (plutôt que d'une distance) permet de raisonner en termes de masse de notes à donner à un système. Ces notes s'additionnent directement lorsque l'on passe d'un fragment d'article individuel à l'ensemble des fragments à dater.
– La note totale S pour un système est la moyenne des notes s_i obtenues pour chaque fragment d'article a_i des N fragments du corpus de test (1) :

$$S = \frac{1}{N} \sum_{i=1}^{N} s\left(d_p\left(a_i\right), d_r\left(a_i\right)\right) \tag{1}$$

Nous choisisons pour calculer la similarité entre date prédite et date de référence la fonction gaussienne (2) :

$$s_g\left(d_p, d_r\right) = e^{-\frac{\pi}{10^2}\left(d_p - d_r\right)^2} \tag{2}$$

Le maximum de s_g vaut 1 pour $d_p = d_r$. La fonction tend vers 0 lorsque d_p s'éloigne de d_r. Le tableau 1 donne les valeurs de s_g en fonction de la valeur absolue de la différence $d_p - d_r$. L'aire sous la courbe (intégrale) de s_g est égale à 10 : la masse totale de score de tolérance offerte à d_p est la même que celle qui serait produite par un intervalle de tolérance de 10 ans centré sur la date de référence d_r et à l'intérieur duquel le score de d_p vaudrait 1 (configuration de DEFT 2010). La fonction s_g remplace le score de similarité binaire de DEFT 2010 (1 si on est dans ces 10 ans, 0 sinon) par une décroissance plus graduelle.

C'est cette fonction de similarité s_g, moyennée sur l'ensemble des N fragments d'articles du corpus (formule 1, précisée en 3), qui a été utilisée pour calculer le score officiel d'un système p dans cette tâche :

$$S\left(p\right) = \frac{1}{N} \sum_{i=1}^{N} e^{-\frac{\pi}{10^2}\left(d_p(a_i) - d_r(a_i)\right)^2} \tag{3}$$

$\lvert d_p - d_r \rvert$	0	1	2	3	4	5	6	7	8
$s_g(d_p, d_r)$	1,000	0,969	0,882	0,754	0,605	0,456	0,323	0,215	0,134
$\lvert d_p - d_r \rvert$	9	10	11	12	13	14	15	> 15	
$s_g(d_p, d_r)$	0,078	0,043	0,022	0,011	0,005	0,002	0,001	0,000	

TAB. 1 – Valeur du score de similarité s_g selon la distance entre deux années. On peut vérifier que la somme de ces valeurs pour $d_p - d_r$ variant entre -15 et $+15$ est 10.

Extension à des hypothèses multiples avec score de confiance Dans la situation où un système donne plusieurs hypothèses de dates pour un fragment d'article, le gain assigné à cet ensemble d'hypothèses est la combinaison linéaire des gains de chaque hypothèse, pondérée par les scores de confiance donnés par le système. Mis en formules : pour un fragment d'article a_i, le système p prédit n_i dates d_p^j :

$$D_p\left(a_i\right) = \left(d_p^1, d_p^2, \ldots, d_p^{n_i}\right)$$

Le système p attribue la confiance c_p^j à la prédiction d_p^j :

$$C_p\left(a_i\right) = \left(c_p^1, c_p^2, \ldots, c_p^{n_i}\right) \text{ avec } \sum_{j=1}^{n_i} c_p^j = 1$$

Le score pondéré obtenu pour ce fragment d'article est alors :

$$s_c\left(a_i\right) = \frac{1}{n_i} C_p\left(a_i\right) \cdot D_p\left(a_i\right) = \frac{1}{n_i} \sum_{j=1}^{n_i} c_p^j \cdot s\left(d_p^j\left(a_i\right), d_r\left(a_i\right)\right) \tag{4}$$

ce qui donne la formule (5) pour l'évaluation d'un fragment d'article avec n_i hypothèses d_p^j pondérées par les scores de confiance c_p^j :

$$s_c\left(a_i\right) = \frac{1}{n_i} \sum_{j=1}^{n_i} c_p^j \cdot e^{-\frac{\pi}{10^2}\left(d_p^j(a_i) - d_r(a_i)\right)^2} \tag{5}$$

et la formule (6) pour l'évaluation globale des résultats d'un système p produisant des hypothèses multiples pondérées par score de confiance :

$$S_c\left(p\right) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} c_p^j \cdot e^{-\frac{\pi}{10^2}\left(d_p^j(a_i) - d_r(a_i)\right)^2} \tag{6}$$

2.3 Tests

2.3.1 Tests humains

Des tests ont été effectués auprès de cinq évaluateurs humains sur les 15 premiers documents du corpus d'apprentissage distribué aux participants. Chaque évaluateur a mis entre 30 et 45 minutes pour travailler sur ce corpus. Les résultats varient fortement : 0,879 – 0,691 – 0,443 – 0,303 et 0,201, pour un score moyen de 0,503. Notons qu'un tirage aléatoire de dates engendre un score final de 0,071. Tous les évaluateurs humains ont effectué un repérage d'entités nommées qu'ils ont ensuite essayé de dater ; ils ont pour la plupart effectué des recherches dans Google et dans l'encyclopédie Wikipédia. Nous observons que l'évaluatrice qui s'est classée première sur ces tests a suivi des études d'histoire et sciences de l'information. L'évaluatrice classée deuxième a, en plus de la datation des entités nommées, également cherché à définir une thématique générale traitée dans le document (conflit, politique, théâtre, etc.) puis effectué une recherche sur Internet combinant cette thématique avec les entités nommées précédemment identifiées.

Afin de représenter la précision de chaque évaluateur et l'évolution des datations, nous avons établi un graphique à la manière des courbes ROC présentant l'incrémentation du nombre de documents correctement identifiés en

augmentant au fur et à mesure la distance en années vis à vis de la référence. Sur ce graphique (voir figure 2), un système précis est celui qui identifiera le maximum de documents avec le minimum d'écart par rapport à la référence, notamment parmi les quinze premières années qui sont celles qui rapportent des points (cf. l'importance des surfaces jaune et vert foncé par rapport aux surfaces vert clair ou bleu clair) : l'évaluatrice classée première a identifié 11 documents avec l'année exacte, lui garantissant une surface importante dès l'année 0 de distance (surface jaune, score de 0,879). À l'inverse, le 11ème document identifié par le dernier évaluateur humain l'a été avec presque 40 ans d'écart (zone bleu clair, score de 0,201). Un tirage aléatoire (zone vert clair, score de 0,071) témoigne d'une précision encore moindre que celle obtenue par les évaluateurs humains. Ces tests révèlent une tâche difficile mais qui devrait engendrer des disparités selon les méthodes suivies par les participants.

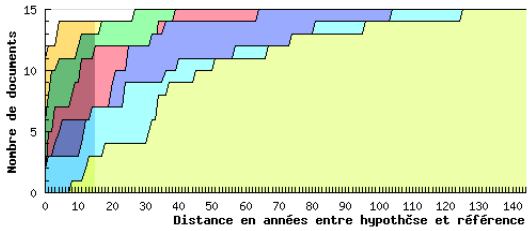

FIG. 2 – Surfaces présentant les résultats des évaluateurs humains sur la tâche diachronie

2.3.2 Test automatique

Nous avons procédé à un deuxième type de test, fondé sur l'adaptation aux données de DEFT2011 d'un système ayant participé à la tâche diachronie pour DEFT2010. Pour ce second test, Rémy Kessler, membre de l'équipe du LIA l'année précédente, a modifié l'un des systèmes entrant dans la composition de la chaîne de traitements utilisée par cette équipe (Oger *et al.*, 2010). Ce système utilise un classifieur à large marge spécialisé dans le traitement de données textuelles, ICSIBOOST[3] développée par le laboratoire ICSI[4] et basé sur un algorithme de boosting (Schapire & Singer, 2000) de classifieurs simples (des arbres de décision à 1 niveau de profondeur sur la présence ou l'absence de n-grammes). L'adaptation de ce système aux données de DEFT2011 s'est faite en l'espace d'une après-midi. L'ensemble des paramètres tels que les prétraitements linguistiques, le remplacement de la ponctuation par des balises lexicales ou T, le nombre de tours de l'algorithme ont été conservés comme lors de l'édition 2010. Les principales modifications ont été l'adaptation des formats d'entrées et de sorties afin de prédire des décennies ou des années en fonction du besoin.

Le système ainsi modifié a été évalué sur le corpus de test de DEFT2011 pour les deux pistes proposées aux participants, avec deux types de sortie (voir tableau 2) : en premier lieu, le système a renvoyé des décennies, à l'image de ce qui était demandé en 2010 ; en second lieu, le système a produit des années, tel que demandé pour la présente édition du défi. L'évaluation sur des décennies est meilleure que celle portant sur des années, ce qui est conforme avec l'idée qu'un nombre plus réduit de classes (15 classes pour les décennies contre 144 classes pour les années) permet d'obtenir de meilleurs résultats. On notera cependant un accroissement important des temps d'apprentissages pour le système avec la version 144 classes. Comparativement aux résultats obtenus par les participants de cette année (voir tableau 3), ce système se classe virtuellement troisième sur les deux pistes, mais avec des scores inférieurs aux moyennes et médianes calculées sur les meilleures soumissions.

	Décennies		**Années**	
Évaluation	Piste 1	Piste 2	Piste 1	Piste 2
Sans confiance	0,236	0,287	0,140	0,167
Avec confiance	—	—	0,109	0,108

TAB. 2 – Scores obtenus par l'adaptation d'un outil DEFT2010 aux données DEFT2011

[3] http://code.google.com/p/icsiboost/
[4] http://www.icsi.berkeley.edu/

2.4 Résultats des participants

Les résultats des participants (tableau 3) sont tout juste inférieurs à la moyenne des tests humains (0,503). Les résultats entre parenthèses correspondent à des soumissions reçues après la fin de la période de test, le participant pensant les avoir soumises en temps et en heure. Nous les indiquons néanmoins car ils représentent un travail conséquent. Comme nous l'envisagions, les résultats sur des extraits de 500 mots sont supérieurs (score moyen de 0,332) à ceux obtenus sur des extraits de 300 mots (score moyen de 0,247). Le corpus provenant d'une reconnaissance optique de caractères sur de journaux anciens, il apparaît légitime d'obtenir de meilleurs résultats sur un ensemble plus vaste de données.

	Piste 1			**Piste 2**		
Équipe et renvoi bibliographique	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
EBSI (Boley, 2011)	0,062	0,073	0,061	0,069	—	—
IRISA (Raymond & Claveau, 2011)	0,342	0,317	**0,472**	0,266	0,285	**0,430**
LIMSI (García-Fernandez *et al.*, 2011)	0,452	0,428	0,363	0,378	0,374	0,358
LUTIN-a (El Ghali, 2011)	(0,098)	(0,108)	(0,100)	0,113	(0,117)	(0,081)
Moyenne	*0,332*			*0,247*		
Médiane	*0,452*			*0,358*		
Écart-type	*0,225*			*0,183*		
Variance	*0,051*			*0,033*		

TAB. 3 – Scores des participants, moyenne, médiane, écart-type et variance sur les meilleures soumissions

Comme pour les tests humains, nous avons représenté la progression des résultats des participants sous la forme de surfaces (figure 3). Puisque les 15 premières années sont les seules à apporter un gain dans l'évaluation finale, nous avons effectué un zoom sur ces premières années (figure 4). Le code couleur utilisé est le suivant (par ordre d'apparition) : IRISA (orange), LIMSI (jaune), LUTIN-a (bleu), EBSI (vert clair).

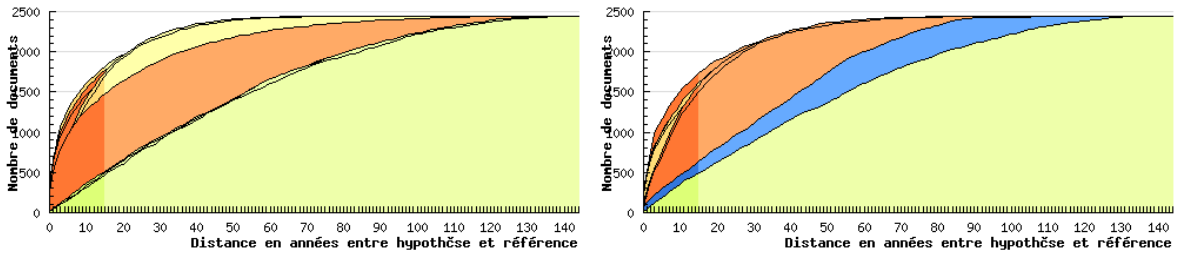

FIG. 3 – Précisions obtenues par les systèmes sur les pistes 1 (graphique de gauche) et 2 (graphique de droite)

2.5 Méthodes des participants

Les participants à cette tâche ont tous eu recours à des méthodes à base d'apprentissage, avec des résultats hétérogènes : un classifieur bayésien naïf pour (Boley, 2011), des SVM par (García-Fernandez *et al.*, 2011), les arbres de décisions et les k-plus-proches voisins par (Raymond & Claveau, 2011). Au niveau linguistique, (Boley, 2011) a comparé les stratégies sémantiques et asémantiques sur ce type de données ; alors que la stratégie sémantique se dégageait nettement sur le corpus d'apprentissage, les résultats sont équivalents sur le corpus de test. (García-Fernandez *et al.*, 2011) ont produit des ressources chronologiques externes aux corpus (une base de données des dates de naissance de personnes célèbres nées entre 1781 et 1944) et effectué une analyse linguistique (étude des archaïsmes et néologismes, études des réformes orthographiques) s'inspirant des travaux de (Albert *et al.*, 2010). Partant du principe que plus un document est ancien, plus la reconnaissance des caractères sera bruitée (les tâches

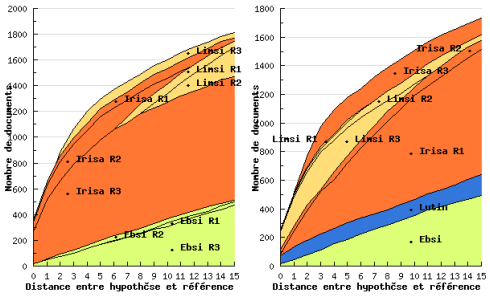

FIG. 4 – Précisions obtenues par les systèmes sur les pistes 1 (graphique de gauche) et 2 (graphique de droite), zoom effectué sur les distances inférieures à 15 ans vis à vis de la référence

d'encre sur un papier fin étant reconnues à tort comme des signes de ponctuation), (Raymond & Claveau, 2011) ont pris en compte la fréquence des ponctuations au fil du temps.

3 Tâche 2. Appariements

Pour cette seconde tâche, nous avons décidé de nous focaliser sur le résumé d'articles scientifiques. Plutôt que de se placer dans une tâche de génération automatique de résumés, contraignante au niveau de l'évaluation – *Qu'est-ce qui constitue un résumé de référence ? Comment évaluer la pertinence d'un résumé ?* – comme c'est le cas dans des campagnes d'évaluation telles TSC sur le résumé automatique (Okumura *et al.*, 2003), nous avons adopté le point de vue inverse qui consiste à utiliser les résumés déjà associés à des articles scientifiques, et à proposer comme tâche d'apparier chaque résumé avec l'article scientifique pour lequel il a été écrit. De récents travaux se sont penchés sur le rapport entre le contenu du résumé et celui de l'article scientifique qui lui correspond, en particulier dans le domaine biomédical (Cohen *et al.*, 2010), permettant de mettre en évidence des caractéristiques propres aux résumés, et propres aux corps des articles scientifiques. Nous émettons l'hypothèse que les méthodes permettant de relier les résumés aux articles devraient permettre de mettre en évidence les éléments saillants d'un résumé, ces éléments et méthodes pouvant par la suite conduire au développement d'outils de génération automatique de résumés.

Dans le cadre de cette tâche, nous avons proposé deux pistes aux participants. En premier lieu, apparier chaque résumé avec l'article scientifique complet qui lui correspond. En second lieu, apparier chaque résumé avec l'article scientifique qui lui correspond, l'article ayant été amputé de son introduction et de sa conclusion. Pour cette seconde piste, nous émettons l'hypothèse que les éléments présents dans l'introduction et la conclusion sont davantage repris dans le résumé, à tout le moins sous une forme quasi identique. Les résultats sur cette piste devraient donc être moins bons que ceux obtenus en étudiant les articles au complet de la première piste.

3.1 Corpus

Les corpus de cette tâche ont été constitués à partir d'articles scientifiques parus dans des revues en Sciences Humaines et Sociales. Ces articles proviennent de la plateforme universitaire Erudit[5]. Chaque article est disponible au format XML avec une structuration permettant la récupération distincte des méta-données (auteurs, résumé, bibliographie, etc.) et du contenu (article scientifique structuré en sections et paragraphes). Nous avons constitué le corpus de la piste 1 en désolidarisant le résumé de l'article de chaque fichier XML, en deux fichiers distincts. Le corpus de la piste 2 a été constitué en supprimant les introduction et conclusion de chaque article scientifique (production d'un fichier de texte scientifique « .txt »). Si les introduction et conclusion n'étaient pas clairement indiquées sous la forme de sections nommées, nous avons considéré les paragraphes précédent le premier titre de section comme étant une introduction, et les paragraphes de la dernière section comme étant la conclusion.

[5] `http://www.erudit.org/` – plateforme de diffusion issue d'un consortium universitaire québécois.

Nous avons utilisé le même jeu d'articles scientifiques pour les deux pistes, les fichiers de résumés, d'articles et de textes ayant été nommés différemment dans les deux pistes.

Le corpus d'apprentissage compte 5 revues (*Anthropologie et Société, Études Internationales, Études littéraires, Philosophiques* et la *Revue des Sciences de l'Éducation*) avec 60 articles par revue. Nous avons ajouté une sixième revue (*Meta*) dans le corpus de test de manière à éprouver la robustesse des systèmes sur une source inconnue ; chaque revue de ce corpus intégrant une trentaine d'articles. Le nombre moyen de mots par article varie selon les revues, avec du nombre moyen le plus long au plus court : *Études Internationales, Revue des Sciences de l'Éducation, Philosophiques, Anthropologie et Société, Études Littéraires,* et enfin *Meta* (voir tableau 4).

Corpus	**Apprentissage**		**Test**	
	Articles	Textes	Articles	Textes
Études Internationales	7044	5557	7032	5236
Revue des Sciences de l'Éducation	6332	5352	6049	5143
Philosophiques	5343	4687	6912	5004
Anthropologie et Société	5549	4358	5889	4540
Études Littéraires	4814	3489	4883	3559
Meta	—	—	4136	3481

TAB. 4 – Nombre moyen de mots par article (pour chaque revue et corpus)

3.2 Méthodes d'évaluation

Nous considérons que l'hypothèse retournée par le système est correcte ou pas (évaluation binaire). On peut compter la proportion de résumés pour lesquels l'hypothèse fournie est correcte. Comme tout résumé doit recevoir une réponse, et que cette réponse est unique, cette proportion peut être vue aussi bien comme une précision (proportion des réponses proposées qui sont correctes) que comme un rappel (proportion des réponses attendues qui sont correctement proposées par le système) ou encore une correction (proportion des décisions qui sont correctes).

Si l'on définit un score élémentaire pour chaque résumé qui vaut 1 ou 0 selon que l'article trouvé est correct ou pas, cela revient à calculer la moyenne de ce score sur l'ensemble des résumés. Mis en formules, pour chacun des n résumés r_i, le système prédit quel article $a_p(r_i)$ parmi les N articles a_j lui correspond. Le score $s(a_p(r_i), a_r(r_i))$ donné à chaque prédiction vaut 0 ou 1 selon que l'article prédit $a_p(r_i)$ est ou pas l'article de référence $a_r(r_i)$:

$$s(a_p(r_i), a_r(r_i)) = \{1 \; si \; a_p(r_i) = a_r(r_i) \; 0 \; sinon$$

Le score global est la moyenne des scores obtenus par le système p :

$$S(p) = \frac{1}{N}\sum_{i=1}^{n} s(a_p(r_i), a_r(r_i)) = \frac{1}{N}\sum_{i=1}^{N} |\{r_i \,;\; a_p(r_i) = a_r(r_i)\}| \tag{7}$$

Extension à des hypothèses multiples avec indice de confiance Dans cette variante, le système peut donner plusieurs hypothèses d'articles pour chaque résumé, en associant un indice de confiance à chaque hypothèse. Si l'une de ces hypothèses est correcte, le score attribué au système est l'indice de confiance que le système a associé à cette hypothèse ; si aucune hypothèse n'est correcte, le score est nul. Comme dans le cas à une seule étiquette, le score global pour l'ensemble des résumés est la moyenne des scores par résumé.

Mis en formules : pour un résumé r_i, le système p prédit n_i étiquettes a_p^j :

$$A_p(r_i) = \left(a_p^1, a_p^2, \ldots, a_p^{n_i}\right)$$

Le système attribue la confiance c_p^j à la prédiction a_p^j :

$$C_p(r_i) = \left(c_p^1, c_p^2, \ldots, c_p^{n_i}\right) \text{ avec } \sum_{j=1}^{n_i} c_p^j = 1$$

Le score pondéré obtenu pour ce résumé est alors :

$$s_c\left(A_p\left(r_i\right), C_p\left(r_i\right), a_r\left(r_i\right)\right) = \{\, c_p^{j_i}\left(r_i\right) si\ \exists j_i \in \{1 \ldots n_i\}\ ;\ a_p^{j_i}\left(r_i\right) = a_r\left(r_i\right) 0 sinon \tag{8}$$

ce qui donne la formule (9) pour l'évaluation globale des résultats d'un système p produisant des hypothèses multiples pondérées par score de confiance :

$$S_c\left(p\right) = \frac{1}{N}\sum_{i=1}^{N} s_c\left(A_p\left(r_i\right), C_p\left(r_i\right), a_r\left(r_i\right)\right) = \frac{1}{N}\sum_{\left\{i\,;\,\exists j_i \in \{1 \ldots n_i\}\,;\,a_p^{j_i}\left(r_i\right) = a_r\left(r_i\right)\right\}} c_p^{j_i}\left(r_i\right) \tag{9}$$

3.3 Tests humains

Les évaluateurs humains ont travaillé à l'identification de 15 couples résumé/article et résumé/texte provenant de deux revues (*Anthropologie et Société* et *Études Internationales*). Ils ont tous correctement identifié les paires résumé/article et résumé/texte, obtenant le score maximal, en au plus d'une demi-heure. Nous en concluons que la tâche est aisée pour un humain et ne devrait pas poser trop de problèmes pour un système.

3.4 Résultats des participants

Pendant la phase de test, un participant nous a signalé l'existence d'un fichier texte vide « 077.txt », ce qui se révèle particulièrement handicapant pour l'aligner avec le fichier de résumé qui lui correspond, « 017.res ». La correspondance a été communiquée à ce participant. Pour les autres participants, nous avons fait le choix de ne pas transmettre cette information mais de modifier le script d'évaluation en conséquence : premièrement, pour ne pas évaluer la sortie sur « 017.res » (appariemment forcément erroné) et deuxièmement, pour accorder gratuitement un point correspondant au point perdu par le participant qui apparie le fichier « 077.txt » avec un résumé quelconque.

3.4.1 Réponse unique : classement officiel

Il s'agit des mesures qui seront utilisées pour le classement final. Si le participant a envoyé des soumissions sans score de confiance, nous prenons la réponse fournie. Si le participant a transmis des soumissions avec score de confiance, nous ne retenons que la réponse ayant le score de confiance le plus élevé.

	Piste 1			**Piste 2**		
Équipe et renvoi bibliographique	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
CHART (Hoareau *et al.*, 2011)	**1,000**	—	—	0,995	—	—
EBSI (Boley, 2011)	0,980	0,985	0,980	0,954	0,954	—
FBK (Tonelli & Pianta, 2011)	0,975	0,960	0,990	0,964	0,934	0,964
GREYC (Lejeune *et al.*, 2011)	**1,000**	0,975	0,626	0,909	0,482	—
INAOE (Sánchez-Vega *et al.*, 2011)	0,970	0,960	0,949	0,904	0,848	0,858
IRISA (Raymond & Claveau, 2011)	0,995	—	—	0,990	—	—
LORIA (Cadot *et al.*, 2011)	**1,000**	—	—	**1,000**	—	—
LUTIN-a (El Ghali, 2011)	0,965	0,934	0,970	0,919	0,883	0,873
LUTIN-d (Devatman Hromada, 2011)	0,909	—	—	0,873	—	—
UCL (Bestgen, 2011)	**1,000**	—	—	**1,000**	—	—
UPF (Saggion, 2011)	0,975	0,975	—	0,959	0,959	—
Moyenne	*0,981*			*0,956*		
Médiane	*0,990*			*0,959*		
Écart-type	*0,027*			*0,042*		
Variance	*0,001*			*0,002*		

TAB. 5 – Résultats des participants sur les réponses de rang 1, moyenne, médiane, écart-type et variance calculés sur les meilleures soumissions

3.4.2 Réponses avec confiance : évaluation alternative

Cette seconde évaluation prend en compte toutes les réponses fournies par le participant et pondère le score d'association par la confiance renseignée par le système. Alors que dans le classement officiel, une réponse juste vaut 1 point, ici elle vaut la valeur du score de confiance qui lui a été associée, ce qui conduit à une dégradation des résultats. Les participants qui n'ont pas utilisé de score de confiance obtiennent les mêmes scores que dans le classement officiel (cela revient à avoir affecté une confiance de 1 à chaque résultat).

	Piste 1			**Piste 2**		
Équipe	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
INAOE	0,417	0,412	0,389	0,368	0,345	0,348
UPF	**0,512**	—	—	**0,402**	—	—
Moyenne	*0,465*			*0,385*		
Médiane	*0,465*			*0,385*		
Écart-type	*0,067*			*0,024*		
Variance	*0,005*			*0,001*		

TAB. 6 – Résultats des participants avec prise en compte du score de confiance, moyenne, médiane, écart-type et variance calculés sur les meilleures soumissions

3.5 Méthodes des participants

La majorité des participants a envisagé cette tâche comme étant de la recherche d'information. (Raymond & Claveau, 2011) ont ainsi considéré que le résumé constituait la requête alors que (Lejeune *et al.*, 2011) ont adopté la démarche inverse consistant à apparier un article avec les différents résumés. La majorité des participants a utilisé une représentation vectorielle des documents, parfois avec une pondération issue du tf*idf et une lemmatisation, la similarité étant généralement calculée au moyen du cosinus ou de la distance euclidienne. (Hoareau *et al.*, 2011) ont combiné les espaces sémantiques vectoriels aux modèles de graphe. (Cadot *et al.*, 2011) ont mis au point une méthode d'appariement qui s'apparente à celle des voisins réciproques « résumé-texte et texte-résumé », en univers fermé, robuste et sans nécessité d'information extérieure. (Bestgen, 2011) a développé une approche fondée sur trois composants : l'analyse sémantique latente (LSA), les machines à support vectoriel (SVM) et l'assignation finale selon l'algorithme du meilleur d'abord (MA). (Devatman Hromada, 2011) a utilisé une approche simple consistant à comparer la fréquence d'utilisation des mots dans les résumés avec celle dans les articles.

Conclusion

La première tâche de datation d'archives de journaux issues d'une reconnaissance de caractères a donné lieu à l'utilisation de différentes méthodes d'apprentissages, certaines étant combinées avec des ressources linguistiques. Contrairement à l'édition 2010 où nous attendions des participants qu'ils identifient la décennie de publication, l'édition 2011 s'est focalisée sur l'année exacte, soit 144 années contre 15 décennies. Les résultats obtenus cette année se révèlent néanmoins plus élevés que ceux obtenus lors de la précédente édition. La méthode ayant obtenu les meilleurs résultats repose sur l'utilisation des k-plus-proches voisins sans recourir à des données externes mais en associant des étiquettes morpho-syntaxiques à chaque token. La seconde tâche relative aux appariements résumés/articles et résumés/textes s'est révélée facile à traiter, au regard des résultats obtenus par les participants. La majorité des méthodes utilisées ne nécessite pas de ressources externes et a été envisagée comme étant une tâche de recherche d'information.

Remerciements

Nous remercions la plateforme Erudit pour la mise à disposition des articles scientifiques et le portail Gallica pour la possibilité d'utiliser les archives de presse. Nous remercions les évaluateurs humains (Marcela Baiocchi,

Roxane Cayer-Tardif, Janie Gauthier-Boudreau et Laurie-Anne Gignac) pour le temps passé à tester la faisabilité des différentes tâches et les résultats obtenus, ces derniers nous ayant confortés dans l'opportunité de proposer ces tâches aux participants. Nous remercions également tous les participants de l'édition 2011 pour les méthodes originales qu'ils ont développées dans le cadre de ce défi et les idées nouvelles qu'ils ont apportées au domaine. Enfin, nous remercions Rémy Kessler, membre de l'équipe du LIA à DEFT2010, pour avoir adapté son système de datation aux données de DEFT2011 de manière à tester la tâche sur la diachronie et à évaluer les possibilités d'adaptation d'un tel système sur de nouvelles données.

Ces travaux ont été en partie réalisés dans le cadre du programme Quaero, financé par Oseo, agence française pour l'innovation.

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 DEFT 2010*.

BESTGEN Y. (2011). LSVMA : au plus deux composants pour apparier des résumés à des articles. In *Actes DEFT 2011*.

BOLEY R. (2011). Comparaison de méthodes sémantiques et asémantiques pour la catégorisation automatique de documents. In *Actes DEFT 2011*.

CADOT M., AUBIN S. & LELU A. (2011). Indexer, comparer, apparier des textes et leurs résumés : une exploration. In *Actes DEFT 2011*.

COHEN K. B., JOHNSON H. L., VERSPOOR K., ROEDER C. & HUNTER L. E. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, **11**(492).

DEVATMAN HROMADA D. (2011). One simple formula for losing DEFT with more than 90% of correct guesses. In *Actes DEFT 2011*.

EL GHALI A. (2011). Expérimentations autour des espaces sémantiques hybrides. In *Actes DEFT 2011*.

GARCÍA-FERNANDEZ A., LIGOZAT A.-L., DINARELLI M. & BERNHARD D. (2011). Méthodes pour l'archéologie linguistique : datation par combinaison d'indices temporels. In *Actes DEFT 2011*.

GROUIN C., FOREST D., SYLVA L. D., PAROUBEK P. & ZWEIGENBAUM P. (2010). Présentation et résultats du défi fouille de texte DEFT2010 : Où et quand un article de presse a-t-il été écrit ? In *Actes DEFT 2010*.

HOAREAU Y. V., AHAT M., PETERMANN C. & BUI M. (2011). Couplage d'espaces sémantiques et de graphes pour le deft 2011 : une approche automatique non supervisée. In *Actes DEFT 2011*.

LEJEUNE G., BRIXTEL R. & GIGUET E. (2011). Deft 2011 : Appariement de résumés et d'articles scientifiques fondé sur des similarités de chaînes de caractères. In *Actes DEFT 2011*.

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 DEFT 2010*.

OKUMURA M., FUKUSIMA T. & NANBA H. (2003). Text summarization challenge 2 : text summarization evaluation at NTCIR workshop 3. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop*.

RAYMOND C. & CLAVEAU V. (2011). Participation de l'IRISA à DEFT2011 : expériences avec des approches d'apprentissage supervisé et non supervisé. In *Actes DEFT 2011*.

SAGGION H. (2011). Matching Texts with SUMMA. In *Actes DEFT 2011*.

SÁNCHEZ-VEGA F., VILLATORO-TELLO E., JUÁREZ-GOZÁLEZ A., VILLASENOR-PINEDA L., Y GÓMEZ M. M. & MENESES-LERÍN L. (2011). INAOE DEFT2011 : Using a Plagiarism Detection Method for Pairing Abstracts-Scientific Papers. In *Actes DEFT 2011*.

SCHAPIRE R. E. & SINGER Y. (2000). BoosTexter : A Boosting-based System for Text Categorization. *Machine Learning*, **39**(2/3), 135–168.

TONELLI S. & PIANTA E. (2011). Matching documents and summaries using key-concepts. In *Actes DEFT 2011*.