Analyse automatique de textes littéraires et scientifiques : présentation et résultats du défi fouille de texte DEFT2014

Thierry Hamon [1,2] Quentin Pleplé [3] Patrick Paroubek [1]
Pierre Zweigenbaum [1] Cyril Grouin [1]
(1) LIMSI-CNRS, Campus universitaire d'Orsay, Rue John von Neumann, Bât 508, 91405 Orsay
(2) Université Paris 13, Villetaneuse, France
(3) ShortEdition, 12 rue Ampère, 38000 Grenoble
prenom.nom@limsi.fr, quentin@short-edition.com

Résumé. Dans cet article, nous présentons l'édition 2014 du défi fouille de texte (DEFT) consacrée à l'analyse de textes littéraires (corpus Short Edition) et scientifiques (archives TALN) au travers de quatre tâches : catégoriser le genre littéraire d'une œuvre, évaluer la qualité littéraire, déterminer l'aspect consensuelle d'une œuvre auprès des relecteurs, et identifier la session d'appartenance d'un article scientifique dans une conférence. Afin d'évaluer les résultats des participants, nous avons utilisé le gain cumulé normalisé (NDCG, tâche 1), l'exactitude en distance relative à la solution moyenne (EDRM, tâche 2), la précision (tâche 3), et la correction (tâcheé4). Les résultats obtenus par les participants sont fortement contrastés et témoignent de la difficulté de chacune des tâches, bien qu'un système ait obtenu une performance maximale dans la tâche 4.

Abstract. In this paper, we present the 2014 DEFT text mining shared task, dedicated to the analysis of literature texts (corpus Short Edition) and scientific texts (TALN archives) through four tasks: identifying the literary type, evaluating writing quality, determining whether the quality of a work achieves consensus among the reviewers, and finally identifying the conference session of a scientific paper. In order to evaluate the results, we used normalized discounted cumulative gain (NDCG, task 1), accuracy of the relative distance to the mean solution (EDRM, task 2), precision (task 3), and correction (task 4). The results obtained by the participants are highly contrasted and reveal the difficulty of each task, although one system reached the maximal performance in task 4.

Mots-clés : Fouille d'opinion, classification automatique, évaluation.

Keywords: Opinion mining, automatic classification, evaluation.

1 Introduction

Dans cette édition du défi fouille de textes, nous proposons quatre tâches d'analyse concernant d'une part des textes littéraires (courte littérature), et d'autre part des articles scientifiques :
- Catégoriser le genre littéraire de courtes nouvelles parmi 30 catégories (poésie, nouvelles, policier, etc.) ;
- Évaluer la qualité littéraire de chacune de ces nouvelles en prédisant la note que donnerait un juge humain ;
- Déterminer, pour chacune des nouvelles, si elle est consensuelle auprès des différents relecteurs ;
- Pour chaque édition précédente de TALN, identifier dans la liste des sessions de chaque conférence celle de chaque articles scientifique présenté en communication orale.

Les participants sont autorisés à utiliser toutes les ressources complémentaires qu'ils souhaitent, à l'exclusion des ressources utilisées par les organisateurs pour servir de base à la constitution des corpus (par ex., les pages des sites Short Edition et Archives TALN) ainsi que tout autre source reproduisant tout ou partie de ces informations telle que sites des conférences ou annonces des programmes, à condition de les mentionner avec leur provenance, lors de la présentation de leurs résultats.

2 Corpus

2.1 Textes littéraires

Le corpus des textes littéraires provient du site Short Edition [1], éditeur en ligne de littérature courte.

Les œuvres publiées sont classées parmi quatre catégories principales (sur la gauche de la figure 1) et plusieurs sous-catégories (sur la droite de la figure 1). Certaines sous-catégories sont spécifiques à une catégorie principale (la catégorie des poèmes dispose de neuf sous-catégories qui lui sont propres), tandis que les autres sous-catégories sont applicables à n'importe quelle catégorie principale, y compris pour la catégorie des poèmes. Chaque œuvre peut appartenir à aucune ou plusieurs sous-catégories (au maximum cinq). Tous les poèmes labellisés ont exactement une des neuf sous-catégories qui leur sont spécifiques et aucune ou jusqu'à cinq sous-catégories non-spécifiques.

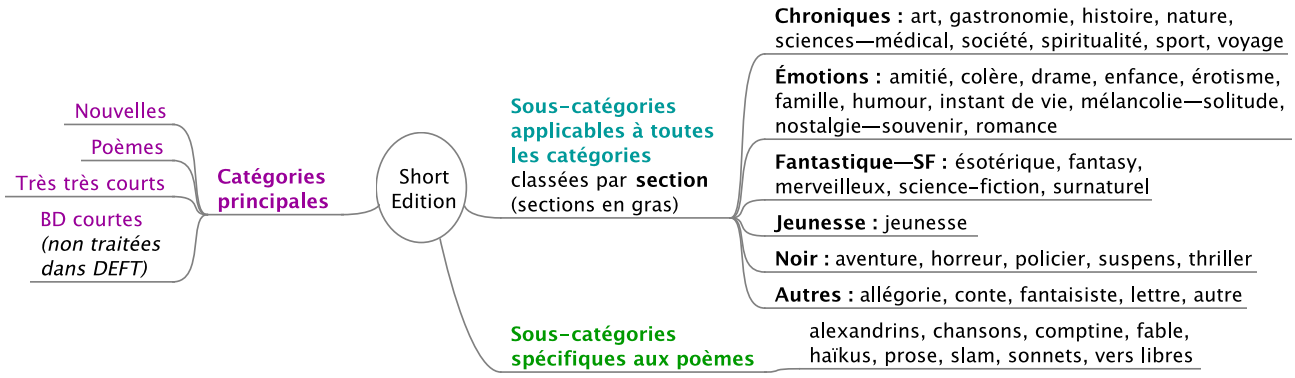

FIGURE 1 – Catégories principales et sous-catégories du système de classification des œuvres du site Short Edition

La classification des œuvres en sous-catégories est établie par quatre personnes chez Short Edition, ce qui confère une cohérence dans la classification opérée. Ces personnes maitrisent la logique des sous-catégories. Ce n'est donc pas l'auteur qui choisit les sous-catégories de l'œuvre qu'il soumet. Les sous-catégorie d'une œuvre sont ordonnées par ordre d'importance : la première est la sous-catégorie principale, etc.

2.2 Textes scientifiques

Le corpus des textes scientifiques est composé des articles parus dans les actes des conférences TALN, disponibles sur le site TALN Archives [2] (Boudin, 2013).

3 Présentation

3.1 Tâches proposées

3.1.1 Corpus de textes littéraires

Tâche 1 – Catégoriser le genre littéraire de courtes nouvelles La première tâche a pour but d'évaluer la capacité d'un système à classer un court texte littéraire selon le genre qui lui correspond. La liste des genres littéraires correspond aux sous-catégories définies par l'éditeur Short Edition. La mise en œuvre de cette classification revêt différents aspects : les aspects stylistiques (vers, mise en forme du texte, etc.), sémantiques (champs sémantiques utilisés, etc.) et syntaxiques.

Nous montrons dans le tableau 1 la répartition des annotations en catégories/sous-catégories sur les corpus d'entraînement et de test de la tâche 1.

1. http://www.short-edition.com/
2. http://www.atala.fr/taln_archives/ ou https://github.com/boudinfl/taln-archives

Sous-catégorie / Section	Corpus	
	Entraînement	Test
instant de vie / émotions	586	237
vers libres / poésie	508	211
société / chronique	440	190
romance / émotions	357	150
drame / émotions	355	147
famille / émotions	290	126
mélancolie–solitude / émotions	279	104
nature / chronique	268	112
nostalgie–souvenirs / émotions	255	95
arts / chronique	200	85
humour / émotions	196	75
enfance / émotions	131	66
fantaisiste / autres	143	57
alexandrins / poésie	140	61
suspens / noir	96	42
histoire / chronique	91	31
amitié / émotions	80	33
voyage / chronique	64	26
conte / autres	59	23
surnaturel / fantastique–SF	54	33
érotisme / émotions	46	27
science-fiction / fantastique–SF	45	26
allégorie / autres	39	26
colère / émotions	39	25
sonnets / poésie	40	24
merveilleux / fantastique–SF	40	13
spiritualité / chronique	40	13
sport / chronique	40	10
sciences–médical / chronique	33	10
prose / poésie	32	15
jeunesse / jeunesse	29	11
chanson / poésie	26	9
aventure / noir	25	13
policier / noir	23	10
comptine / poésie	20	8
thriller / noir	17	1
gastronomie / chronique	14	10
horreur / noir	14	5
slam / poésie	14	3
fable / poésie	13	8
lettre / autres	9	5
autres / autres	6	3
fantasy / fantastique–SF	4	6
haïkus / poésie	3	3
ésotérique / fantastique–SF	3	1
Total	5182	2189

TABLE 1 – Répartition des annotations en catégories/sous-catégories sur les corpus d'entraînement et de test de la tâche 1. Les catégories sous les pointillés présentent des annotations avec un pourcentage inférieur à 1% du nombre total d'annotations dans le corpus

Tâche 2 – Évaluer la qualité littéraire La deuxième tâche propose d'évaluer la qualité littéraire de chacun de ces textes en prédisant la note attribuée par le comité de relecture à chacun des textes littéraires. La référence de cette tâche

est constituée par l'ensemble des notes attribuées par le comité de relecture de l'éditeur Short Edition. Ces notes seront fournies avec le corpus d'entraînement (de 1 « excellent » à 5 « très mauvais »). Une sixième valeur, restée présente dans les corpus distribués, ne renvoie pas à l'évaluation de la qualité de l'œuvre, mais détermine le statut « hors ligne éditoriale » de l'œuvre. Les relectures associées à cette sixième valeur n'ont pas été prises en compte lors de l'évaluation, comme cela a été indiqué aux participants au début de la phase de tests.

Nous représentons sur la figure 2 la répartition des notes attribuées par les relecteurs dans chaque valeur possible, pour les corpus d'entraînement et de test de la tâche 2. On observe une répartition similaire des notes dans les deux corpus, avec une prévalence importante pour les notes 3 à 5 qui constituent l'essentiel des notes attribuées par les relecteurs. Inversement, la note 1 correspondant à une œuvre excellente est utilisée dans 1% du nombre total de relectures seulement. Enfin, les œuvre qualifiées de « hors ligne éditoriale » apparaissent dans 2,7 à 2,8% du nombre total de relectures.

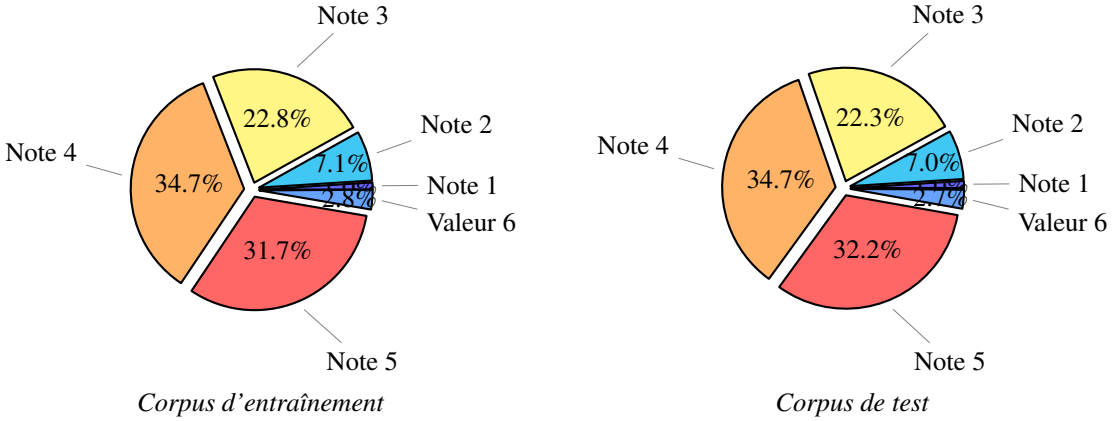

FIGURE 2 – Répartition des notes attribuées par les relecteurs dans les corpus de la tâche 2 (les notes 1 à 5 renvoient à la qualité littéraire, la valeur 6 désigne une œuvre hors ligne éditoriale)

Tâche 3 – Déterminer si une œuvre fait consensus La troisième tâche consiste à déterminer si la qualité d'un texte littéraire fait consensus auprès des différents membres du comité de relecture. La distribution des notes attribuées à chaque œuvre sera fournie avec le corpus d'entraînement. Une œuvre est jugée consensuelle si les notes attribuées par les différents relecteurs ne varient pas au-delà d'un écart de 1 point entre les différentes relectures associées à cette œuvre.

Nous représentons sur la figure 3 la répartition des œuvres selon qu'un consensus entre relecteurs a été observé ou non dans les corpus d'entraînement et de test.

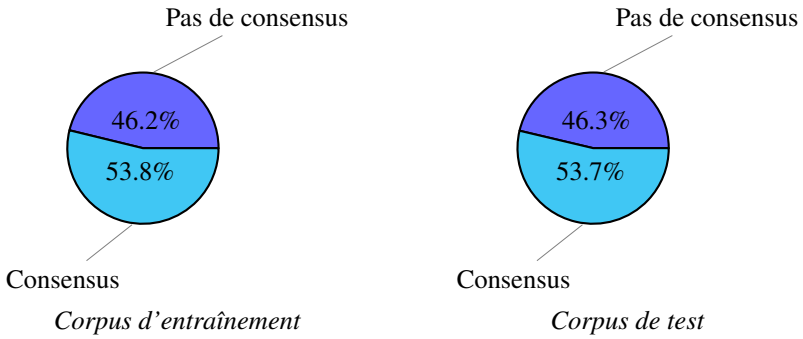

FIGURE 3 – Répartition des œuvres selon qu'un consensus entre relecteurs a été observé ou non (tâche 3)

3.1.2 Corpus de textes scientifiques

Tâche 4 – Déterminer la session scientifique dans laquelle un article de conférence a été présenté La quatrième tâche se démarque des précédentes car elle concerne les articles scientifiques présentés lors des dernières conférences

TALN. Le corpus se composera des articles présentés en communication orale (ni poster, ni conférence invitée). Pour chaque édition, seront fournis : un ensemble d'articles (titre, résumé, mots-clés, texte), la liste des sessions scientifiques de cette édition, et la correspondance article/session (sauf pour le test). Le corpus de test se composera d'une édition complète de TALN (articles et liste des sessions) pour laquelle il faudra identifier dans quelle session chaque article a été présenté.

Des noms de sessions absentes du corpus d'apprentissage peuvent exister dans le corpus de test. Cependant, les listes des sessions utilisées chaque année seront fournies à l'appui du corpus de test, comme elles le sont déjà pour le corpus d'apprentissage.

3.2 Tests humains

Corpus de textes scientifiques Des tests humains ont été réalisés sur la quatrième tâche auprès des étudiants de la promotion 2013/2014 du M2 professionnel d'ingénierie linguistique de l'INaLCO [3]. Les étudiants ont reçu pour consigne d'étudier rapidement le contenu de chaque article (titre, résumé, mots-clés, contenu) et de déterminer la session scientifique la plus probable sous laquelle chaque article a été présenté en conférence. Deux précisions ont été apportées : (i) un article ne dépend que d'une seule session scientifique, et (ii) plusieurs articles peuvent appartenir à la même session scientifique, y compris dans le corpus fourni en test.

Deux corpus de dix articles longs chacun [4], parus entre 2008 et 2013, ont été proposés à deux groupes d'étudiants, accompagnés de la terminologie des sessions scientifiques de l'ensemble des éditions TALN (soit une soixantaine de titres de sessions). Chaque corpus comprenait des articles relevant de quatre sessions seulement (sans que les étudiants ne soient informés du nombre total de sessions différentes, ni du nombre maximum d'articles par session dans le corpus), avec dans chacun des deux corpus la même répartition des articles dans les quatre sessions :
– dialogue homme-machine : 1 article par corpus ;
– morphologie et segmentation : 4 articles par corpus ;
– résumé automatique : 1 article par corpus ;
– traduction et alignement : 4 articles par corpus.
Les résultats ont été évalués en termes de score strict (la session a été retrouvée à l'identique entre l'hypothèse et la référence) et de score souple (la session de l'hypothèse est comprise dans la session de référence, plus générique ou regroupant plusieurs sessions). Pour la session de référence « morphologie et segmentation », si la session fournie est « segmentation », parce qu'elle est comprise dans la session plus générique qui l'englobe, un point sera compté dans le score souple, aucun point dans le score strict.

Les scores calculés sur les tests humains sont de :
– score strict : 2,82 en moyenne (le nombre de sessions correctement identifiées à l'identique varie de 2 à 3 sur dix selon le sous-groupe d'étudiants), soit 28,2% de sessions identifiées à l'identique en moyenne ;
– score souple : 6,64 en moyenne (le nombre de sessions identifiées partiellement varie de 4 à 8 sur dix selon le sous-groupe d'étudiants), soit 66,4% de sessions identifiées de manière partielle en moyenne.
Ces faibles scores s'expliquent pour plusieurs raisons : (i) face à dix documents, il est difficile pour un humain de considérer qu'une même catégorie s'applique à quatre documents alors qu'il existe une soixantaine de catégories disponibles, à plus forte raison deux fois de suite (deux sessions de quatre articles), (ii) le choix des sessions est déterminés par les organisateurs des conférences, certains choix pouvant être dictés par des considérations organisationnelles (contraintes de planning) plutôt que scientifiques, et (iii) les étudiants n'ont pas encore acquis l'habitude des conférences et des articles scientifiques.

Corpus de textes littéraires En raison des contraintes de confidentialité qui pèsent sur le corpus de textes littéraires, nous n'avons pas fait travailler les étudiants sur les données du site Short Edition.

3. Ces tests ont été réalisés dans le cadre du cours de fouille de texte assuré par Cyril Grouin auprès d'un groupe de quinze étudiants des parcours « Ingénierie multilingue » et « Traductiques et gestion de l'information » : Florence BARBEROUSSE, Amélie BOSC, Qinran DANG, Loïc DUMONET, Lucie GIANOLA, Ching Wen HUANG, Guillaume DE LAGANE DE MALÉZIEUX, Jennifer LEWIS WONG, Yingying MA, Amélie MARTIN, Dalia MEGAHED, Satenik MKHITARYAN, Fatemeh SAJADI ANSARI, Phuong Thao TRAN THI, Li Yun YAN. Nous remercions ces étudiants pour le travail qu'ils ont accompli en jouant le rôle d'évaluateurs humains de la tâche.

4. Le premier corpus comprend les articles suivants (les identifiants sont ceux du site TALN Archives) : taln-2008-long-008, taln-2008-long-010, taln-2011-long-022, taln-2011-long-024, taln-2011-long-036, taln-2011-long-038, taln-2013-long-018, taln-2013-long-024, taln-2013-long-030 et taln-2013-long-032. Le deuxième corpus comprend les articles taln-2008-long-009, taln-2008-long-011, taln-2011-long-021, taln-2011-long-023, taln-2011-long-025, taln-2011-long-037, taln-2013-long-001, taln-2013-long-006, taln-2013-long-023 et taln-2013-long-028.

3.3 Organisation

Calendrier Les inscriptions ont été ouvertes le 17 février 2014. Les données d'entraînement ont été distribuées à partir du 12 mars aux équipes ayant complété et signé les contrats d'accès aux données. Les données de test ont été communiquées entre le 12 et le 18 mai, la phase de test étant comprise dans une période de trois jours au libre choix de chaque équipe (accès aux données de test le premier jour, soumission des fichiers de résultats avant la fin du troisième jour). L'atelier de clôture s'est tenu le 1 er juillet, pendant la conférence TALN/RECITAL 2014 à Marseille.

Participants Dix-sept équipes se sont inscrites. Quinze équipes ont accédé aux données d'entraînement, et neuf équipes ont accédé aux données de test. Au terme du défi, sept équipes ont soumis des fichiers de résultats, par ordre alphabétique des affiliations :
– GREYC, Caen (14) : Charlotte Lecluze et Gaël Lejeune ;
– IRIT, Toulouse (31), LIMSI, Orsay (91), LLF, Paris (75) : Farah Benamara, Véronique Moriceau et Yvette Yannick Mathieu ;
– LIA, Avignon (84), ADOC Talent Management, Paris (75) : Luis Adrián Cabrera-Diego, Stéphane Huet, Bassam Jabaian, Alejandro Molina, Juan-Manuel Torres-Moreno, Marc El Bèze et Barthélémy Durette ;
– LIMSI, Orsay (91) : Eva D'hondt ;
– LINA, Nantes (44), IRISA, Rennes (35), LIPN, Villetaneuse (93) : Solen Quiniou, Peggy Cellier et Thierry Charnois ;
– Lutin UserLab, Paris (75) : Adil El Ghali et Kaoutar El Ghali ;
– ÚRK, Bratislava, Slovaquie, CHArt, Saint-Denis (93) : Daniel Devatman Hromada.

4 Méthodes des participants

Tâche 1 – Catégoriser le genre littéraire de courtes nouvelles Pour cette première tâche, tous les participants ont effectué une analyse stylistique des œuvres littéraires, pour en dégager des propriétés relatives à chaque catégorie. Ces propriétés ont ensuite été utilisées, soit directement, soit dans le cadre d'un apprentissage statistique.

Sur cette tâche, (Lecluze & Lejeune, 2014) ont considéré que le style littéraire des œuvres détermine la catégorie littéraire d'appartenance. Les auteurs ont également pris en compte les éléments du texte appartenant à divers champs lexicaux, constatant un bénéfice dans la classification. Cette approche globale semble pertinente au vu des résultats obtenus par l'équipe (voir tableau 2). De manière similaire, (D'hondt, 2014) a identifié dans les documents des indices stylistiques, syntaxiques, et les éléments du texte appartenant à un lexique d'opinions. Ces indices ont ensuite été utilisés pour construire un modèle par apprentissage au moyen d'un perceptron, en limitant le nombre de prédictions à 1 à 5 catégories par document traité. Enfin, l'approche utilisée par (El Ghali & El Ghali, 2014) repose sur les espaces sémantiques avec prise en compte des caractéristiques stylistiques des œuvres. D'autre part, l'approche retenue repose également sur l'utilisation d'arbres de décision pour chaque genre poétique (alexandrin, haïku, prose, sonnet, etc.).

Tâche 2 – Évaluer la qualité littéraire Afin d'évaluer la qualité littéraire des œuvres, (Benamara *et al.*, 2014) ont projeté un lexique d'opinions sur chaque mot des documents, dont les valeurs ont ensuite été utilisées comme caractéristiques pour construire un modèle par apprentissage statistique au moyen d'une régression logistique. Cette approche hybride a permis à cette équipe d'obtenir les meilleurs résultats (voir tableau 3). L'approche suivie par (Lecluze & Lejeune, 2014) repose sur l'identification de motifs récurrents porteurs d'opinion présents dans les relectures pour évaluer la qualité globale de chaque œuvre.

Tâche 3 – Déterminer si une œuvre fait consensus Sur cette tâche, (Benamara *et al.*, 2014) ont utilisé la même approche que celle suivie sur la tâche 2, avec l'obtention de bons résultats (voir tableau 4), tandis que (Lecluze & Lejeune, 2014) ont réutilisé l'approche d'analyse du style littéraire utilisée sur la tâche 1.

Tâche 4 – Déterminer la session scientifique dans laquelle un article de conférence a été présenté Sur cette dernière tâche, (El Ghali & El Ghali, 2014) ont considéré le problème sous l'angle d'un *clustering* au moyen de l'outil K-means. L'approche suivie repose sur la définition de clusters dont le barycentre est modifié jusqu'à aboutir à une convergence, ainsi que de deux types de contraintes : le nombre maximum d'articles par session et des distances entre documents avec

des coûts associés de violation des contraintes. Cette approche globale a permis de correctement simuler la répartition des articles en sessions, l'équipe obtenant un score parfait (voir tableau 5). Une approche différente a été suivie par (Cabrera-Diego *et al.*, 2014) qui ont réalisé plusieurs systèmes dont ils ont combinés et optimisés les résultats : un système collégial combinant plusieurs approches (cosinus, n-grammes, modèle de Poisson, similarité de type Jaccard, k plus proches voisins) utilisées dans le cadre d'une validation croisée, un système reposant sur la similarité de profils, et un système à base de CRF. De manière plus basique, (Lecluze & Lejeune, 2014) ont appliqué une approche consistant à rechercher dans les articles les termes présents dans les noms de session scientifique, partant du principe que les termes utilisés dans les articles se reflètent dans les noms de session. L'approche suivie par (Quiniou *et al.*, 2014) consiste à étudier les motifs fréquents identifiés dans les articles qui sont représentés sous la forme de graphes. Des regroupements de graphes similaires ont ensuite été produits pour rassembler les articles et déterminer la session scientifique d'appartenance. Enfin, (Hromada, 2014) a mobilisé une approche à base de vecteurs sémantiques fondée sur les unigrammes et bigrammes de mots présents dans les titres, noms des auteurs, mots-clés et résumés d'articles.

5 Évaluation

Tâche 1 – Catégoriser le genre littéraire de courtes nouvelles La tâche a pour objectif d'identifier les différentes sous-catégories définissant le genre littéraire des nouvelles mais aussi d'ordonner ces sous-catégories suivant leur degré de pertinence. Il est donc nécessaire d'utiliser une mesure d'évaluation qui prennent en compte des réponses multi-catégories et le rang attribué à chaque catégorie. Ainsi, nous avons retenu le gain cumulé normalisé (*Normalized Discounted Cumulated Gain*, NDCG) (Järvelin & Kekäläinen, 2002). Le gain cumulé atténué par le rang (DCG) est défini de la manière suivante pour le document d :

$$DCG_d = \sum_{i=1}^{d} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

rel_i étant le poids de la sous-catégorie i. Le DCG_d est ensuite normalisé par rapport à celui de la liste de référence pour le document d ($IDCG_d$) :

$$nDCG_d = \frac{DCG_d}{IDCG_d}$$

On prend ensuite la moyenne des $nDCG_d$.

Tâche 2 – Évaluer la qualité littéraire La référence de cette tâche est une note correspondant à chaque relecture. Les systèmes participants devaient renvoyer le même type d'information. Nous considérons ensuite la médiane des notes de relectures associées à l'œuvre comme valeur de référence. L'utilisation de la médiane permet d'agréger les valeurs en éliminant les cas extrêmes. Nous avons ensuite évalué les réponses des systèmes en utilisant l'exactitude en distance relative moyenne à la solution (EDRM) que nous avions déjà utilisé lors de l'édition 2013 (Grouin *et al.*, 2013) :

$$EDRM = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{d(s_i, r_i)}{d_{max}(s_i, r_i)}\right) \quad (1)$$

Ainsi, lors de l'évaluation des réponses d'un système s_i, il est important de prendre en compte la valeur absolue de la distance à la référence r_i : $d(s_i, r_i)$. Par exemple, une distance de 1 à la référence doit être moins pénalisante qu'une réponse de 4. Cette distance doit également tenir compte de la distance maximale possible $d_{max}(s_i, r_i)$ en valeur absolue. La distance entre une réponse du système et la référence est ensuite normalisée. L'EDRM est alors calculée en fonction des distances obtenues pour le N œuvres.

Tâche 3 – Déterminer si une œuvre fait consensus Afin d'évaluer les réponses d'un système détectant le consensus d'une œuvre, nous avons utilisé la précision :

$$precision = \frac{TP}{TP + FP} \quad (2)$$

Tâche 4 – Déterminer la session scientifique dans laquelle un article de conférence a été présenté L'objectif de la tâche étant d'associer un article à une session scientifique, nous avons retenu la correction :

$$correction = \frac{|\{a_j | \exists S_i, a_j \in S_i\}|}{\sum_i |S_i|} \quad (3)$$

où a_j est un article bien rangé dans une session S_i et $|S_i|$ le nombre d'articles à ranger dans la sessions S_i. Cette mesure permet d'évaluer globalement la qualité d'affectation des articles.

Le NDCG (tâche 1) et la précision (tâche 3) sont calculés à l'aide du programme `trec_eval` [5], tandis que pour l'EDRM (tâche 2) et la correction (tâche 4), nous avons implémenté nous-mêmes ces mesures d'évaluation.

6 Résultats

Dans cette section, nous renseignons des résultats globaux obtenus par les participants sur chacune des quatre tâches, pour chacune des soumissions effectuées par les équipes. Le classement officiel (top 3) repose sur la meilleure soumission de chaque équipe (valeur en gras). Les tableaux regroupent les différentes soumissions des participants en blocs, ces blocs étant ensuite classés par ordre décroissant du meilleur score obtenu par l'équipe.

Le table 2 donne les résultats officiels des participants sur la tâche de catégorisation des œuvres (tâche 1), classés par ordre décroissant du meilleur score par équipe, avec indication du rang dans le classement final. Les meilleurs résultats de chaque équipe sont mis en gras. Sur cette tâche, la moyenne s'établit à 0,4475, la médiane à 0,4278 et l'écart-type est de 0,0695.

Équipe	GREYC		Lutin		LIMSI		
Soumission	1	2	1	2	1	2	3
NDCG	0,5130	**0,5248**	**0,4278**	0,2599	0,3817	0,3800	**0,3900**
Rang officiel	–	#1	#2	–	–	–	#3

TABLE 2 – Résultats des participants sur la tâche 1

Le tableau 3 donne les résultats officiels des participants sur la tâche de prédiction des notes des relecteurs (tâche 2), classés par ordre décroissant du meilleur score par équipe, avec indication du rang dans le classement final. Les meilleurs résultats de chaque équipe sont mis en gras. Sur cette tâche, la moyenne et la médiane s'établissent à 0,6121 et l'écart-type est de 0,3035.

Équipe	IRIT/LIMSI/LLF			GREYC
Soumission	1	2	3	1
EDRM (sur la médiane)	0,8193	0,8218	**0,8267**	**0,3975**
Rang officiel	–	–	#1	#2

TABLE 3 – Résultats des participants sur la tâche 2

Le tableau 4 donne les résultats officiels des participants sur la tâche de détermination du caractère consensuel d'une œuvre par les différents relecteurs (tâche 3), classés par ordre décroissant du meilleur score par équipe, avec indication du rang dans le classement final. Les meilleurs résultats de chaque équipe sont mis en gras. Sur cette tâche, la moyenne et la médiane s'établissent à 0,5125 et l'écart-type est de 0,1907.

Le tableau 5 donne les résultats officiels des participants sur la tâche d'identification des sessions scientifiques des articles TALN (tâche 4), classés par ordre décroissant du meilleur score par équipe, avec indication du rang dans le classement final. Les meilleurs résultats de chaque équipe sont mis en gras. Sur cette tâche, la moyenne s'établit à 0,5926, la médiane à 0,4815 et l'écart-type est de 0,2860.

5. `http://trec.nist.gov/trec_eval`

Équipe	IRIT/LIMSI/LLF			GREYC
Soumission	1	2	3	1
Précision	0,6453	**0,6473**	0,6401	**0,3776**
Rang officiel	–	#1	–	#2

TABLE 4 – Résultats des participants sur la tâche 3

Équipe	Lutin	LIA			GREYC			LINA/IRISA/LIPN			ÚRK/CHArt		
Soumission	1	1	2	3	1	2	3	1	2	3	1	2	3
Précision	**1,0000**	**0,7593**	0,3704	0,7037	0,4259	**0,4815**	0,4444	0,4259	0,4259	**0,4444**	**0,2778**	0,2222	**0,2778**
Rang officiel	#1	#2	–	–	–	#3	–	–	–	#4	#5	–	–

TABLE 5 – Résultats des participants sur la tâche 4

7 Conclusion

L'édition 2014 du défi fouille de texte (DEFT) a porté sur l'analyse de textes littéraires et scientifiques.

Sur la tâche de catégorisation des œuvres littéraires, les participants ont tenu compte des aspects stylistiques des documents ainsi que des éléments appartenant à certains champs sémantiques pour déterminer la catégorie d'appartenance. La meilleure équipe a obtenu un gain cumulé normalisé (NDCG) de 0,5248.

Sur la tâche d'évaluation de la qualité littéraire de ces œuvres, avec pour référence les notes attribués par les relecteurs professionnels, les participants ont utilisés des ressources pour la fouille d'opinion, soit des lexiques utilisés dans des approches par apprentissage, soit l'identification de motifs récurrents. La meilleure équipe a obtenu une exactitude en distance relative à la solution moyenne (EDRM) de 0,8267.

Sur la tâche de détermination du caractère consensuel d'une œuvre, les participants se sont fondés, soit sur l'étude stylistiques des documents, soit sur la prise en compte des opinions exprimées dans les documents. La meilleure équipe a obtenu une précision de 0,6473.

Enfin, sur la tâche d'identification de la session scientifique pendant laquelle un article scientifique a été présenté pendant les conférences TALN, les participants ont utilisé des approches par apprentissage statistique, notamment en combinant et fusionnant plusieurs systèmes. La meilleure équipe a obtenu une correction parfaite de 1, prédisant exactement le classement réalisé par les humains lors des conférences utilisées pour le jeu de test. Les prédictions réalisées par les participants sur cette tâche ont donné lieu à des résultats fortement contrastés.

Références

BENAMARA F., MORICEAU V. & MATHIEU Y. Y. (2014). Catégorisation sémantique fine des expressions d'opinion pour la détection de consensus. In *Actes de DEFT*, Marseille, France.

BOUDIN F. (2013). TALN archives : une archive numérique francophone des articles de recherche en traitement automatique de la langue. In *Actes de TALN 2013 (Traitement automatique des langues naturelles)*, p. 507–514, Les Sables-d'Olonne : ATALA LINA-LIUM.

CABRERA-DIEGO L. A., HUET S., JABAIAN B., MOLINA A., TORRES-MORENO J.-M., EL-BÈZE M. & DURETTE B. (2014). Algorithmes de classification et d'optimisation : participation du LIA/ADOC à DEFT'14. In *Actes de DEFT*, Marseille, France.

D'HONDT E. (2014). Genre classification using balanced winnow in the DEFT 2014 challenge. In *Actes de DEFT*, Marseille, France.

EL GHALI A. & EL GHALI K. (2014). Combiner espaces sémantiques, structure et contraintes. In *Actes de DEFT*, Marseille, France.

GROUIN C., PAROUBEK P. & ZWEIGENBAUM P. (2013). DEFT2013 se met à table : présentation du défi et résultats. In *Actes de DEFT*, Les Sables-d'Olonne, France.

HROMADA D. D. (2014). Introductory experiments with evolutionary optimization of reflective semantic vector spaces. In *Actes de DEFT*, Marseille, France.

JÄRVELIN K. & KEKÄLÄINEN J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, **20**(4), 422–446.

LECLUZE C. & LEJEUNE G. (2014). DEFT2014, analyse automatique de textes littéraires et scientifiques en langue française. In *Actes de DEFT*, Marseille, France.

QUINIOU S., CELLIER P. & CHARNOIS T. (2014). Fouille de données pour associer des noms de sessions aux articles scientifiques. In *Actes de DEFT*, Marseille, France.