Actes du quatrième défi fouille de texte, DEFT2008, Avignon, France, 13 juin 2008.
Proceedings of the Fourth DEFT Workshop, DEFT2008, Avignon, France, 13th June 2008.
Pages 3-12

TALN 2008, Avignon, 9–13 juin 2008

Présentation de DEFT'08 (DÉfi Fouille de Textes)

Les membres du Comité d'Organisation de DEFT'08 :
Cyril Grouin[1] Jean-Baptiste Berthelin[1] Sarra El Ayari[1]
Martine Hurault-Plantet[1] Sylvain Loiseau[1]
(1) LIMSI-CNRS – BP133 – 91403 Orsay Cedex
{cyril.grouin, jean-baptiste.berthelin, sarra.elayari,
martine.hurault-plantet, sylvain.loiseau}@limsi.fr

Résumé. Dans le cadre de la campagne d'évaluation annuelle DEFT *(défi fouille de textes)*, la quatrième édition a pour objet l'identification de catégories textuelles en genre et en thème. Nous avons utilisé des articles provenant de deux sources, *Le Monde* et Wikipédia, chaque article ayant été rattaché à l'une des neuf catégories extraites de ces corpus. Cet article présente l'objectif de la tâche, les corpus utilisés ainsi que les prétraitements effectués sur ces corpus. Nous reviendrons également sur les tests manuels que nous avons réalisés pour mesurer la faisabilité de la tâche. Enfin, nous détaillerons les mesures utilisées pour évaluer les résultats des participants.

Abstract. Within the annual DEFT evaluation campaign (where DEFT is for "Défi Fouille de Texte"), the fourth edition involves the identification of textual categories in genre and thema. We used articles from two sources : *Le Monde* (a French daily newspaper), and Wikipedia. Each of these articles is connected with one of the nine categories we extracted from both corpora. This paper describes the aim of the task, the corpora that are used, and the preprocessing that has been applied to them. We also examine the way in which we performed manual tests in order to measure the feasibility of the task. Finally, we give the detail of the measurements that served to evaluate the competitors' results.

Mots-clés : Campagne d'évaluation, fouille de textes, catégorisation, genre, thème, indices de confiance.

Keywords: Evaluation campaign, text mining, categorization, genre, confidence indicators.

1 Introduction

L'édition 2008 de DEFT[1] s'inscrit dans le cadre de la conférence JEP/TALN organisée en Avignon du 9 au 13 juin 2008. Le thème retenu cette année concerne l'identification du genre d'un document parmi deux possibilités (journalistique et encyclopédique) et la catégorisation de chaque document parmi neuf catégories thématiques.

Les participants devaient effectuer deux tâches distinctes :
– Tâche n° 1 : identification du genre et du thème (parmi un choix de quatre thèmes) ;
– Tâche n° 2 : identification du thème uniquement (parmi un choix de cinq thèmes, distincts de ceux utilisés dans la tâche 1) mais portant néanmoins sur les deux genres de documents rassemblés pour ce défi.

2 Présentation des corpus

Les corpus ont été constitués à partir de deux sources distinctes : *Le Monde* et Wikipédia en français, constituant ainsi les deux genres de textes à identifier. Par genre, nous entendons des contextes d'écriture différents, en l'occurrence un contexte journalistique (*Le Monde*) et un contexte d'encyclopédie collaborative sur l'Internet (Wikipédia).

Pour chacun de ces corpus, nous avons relevé neuf catégories communes (rubrique dans laquelle a paru un article du *Monde* ou catégorie sous laquelle a été classé un article de Wikipédia) dont nous donnons ci-dessous des représentants de chaque corpus. Par catégorie, nous entendons un ensemble d'articles traitant de la même thématique.

– Art : articles consacrés à l'art et à la culture (danse, peinture, sculpture, théâtre) ;
– Économie : articles consacrés à l'économie et aux entreprises ;
– France : articles de politique nationale française ;
– International : articles de politique internationale ou nationale (sauf politique française) ;
– Littérature : articles relatifs aux livres (critiques, parutions) et à la littérature ;
– Sciences : articles consacrés aux sciences ;
– Société : articles consacrés aux problèmes de société ne relevant pas du domaine politique ;
– Sports : articles traitant du sport (rencontres, résultats, personnalités) ;
– Télévision : articles consacrés à la radio et à la télévision (programmes, fonctionnement).

3 Préparation des données

3.1 Corpus « Le Monde »

Le corpus du journal *Le Monde* nous a été fourni par la société ELDA[2]. Chaque article est enregistré dans un fichier distinct au format XML qui reproduit la structure organisationnelle du journal en distinguant les éléments constitutifs de l'article (titraille, chapô, texte de l'article)

[1] `http://deft08.limsi.fr/`
[2] ELDA : Evaluations and Language resources Distribution Agency, `www.elda.org`

des méta-données associées (date de publication, secteur de rédaction, éléments d'indexation). Le secteur de rédaction nous fournit la catégorie thématique de l'article.

Afin de disposer d'un nombre équivalent d'articles dans chaque catégorie entre les corpus du *Monde* et de Wikipédia, nous n'avons utilisé qu'une sous-partie du corpus d'origine : tous les articles de l'année 2004 pour les catégories *Art, Économie, France, International, Société* et *Télévision*, les articles des années 2004 et 2005 pour les catégories *Littérature* et *Sports*, et l'ensemble des articles de la période 2003 à 2006 pour la catégorie *Sciences*.

Dans un premier temps, nous avons listé tous les articles publiés sous l'une des neuf catégories prédéfinies (voir section 2) en nous fondant sur le secteur de rédaction renseigné dans chaque fichier. Nous avons ensuite converti au format texte brut les articles de cette liste faisant plus de 300 caractères en effaçant les mentions explicites au journal (indications de copyright et références à un article paru dans une édition antérieure du quotidien). Seuls les articles détaillant les résultats du Loto sportif ont été éliminés de ce nouvel ensemble.

3.2 Corpus « Wikipédia »

La constitution du corpus d'articles de Wikipédia a été réalisée à partir de la version complète de la base, datée d'octobre 2007.

Extraction des articles Une différence importante entre le système de catégorisation en rubriques du *Monde* et le système de catégorisation de Wikipédia est que le premier résulte en une partition des articles alors que dans le second, un article peut appartenir à plusieurs catégories à la fois. Nous avons donc utilisé plusieurs stratégies pour extraire de Wikipédia un ensemble d'articles qui soit une partition en catégories.

Dans un premier temps, nous avons représenté l'ensemble des catégories de Wikipédia sous la forme d'un graphe qui organise les catégories en sous-catégories et super-catégories (voir graphe 1). Afin de collecter les articles, nous sommes partis des catégories les plus représentatives de chacune des neuf thématiques définies pour le défi (par exemple, *Économie* et *Société* sur le graphe). À ce niveau, nous constatons qu'il n'existe que peu d'articles directement rattachés à ces catégories racines.

À partir de ces catégories racines, nous avons parcouru le graphe et collecté tous les articles situés sous ces catégories racines ou sous l'une de leurs sous-catégories, en n'allant pas plus loin que trois niveaux de sous-catégories. Nous avons en effet constaté qu'en s'éloignant de la catégorie racine, d'une part, la spécification thématique décroît (notamment par le biais d'une sur-spécialisation de l'article), et d'autre part, le nombre de connexions vers des catégories hétérogènes augmente (et renforce le risque qu'un article dépende de deux catégories racines).

Pour nous assurer que chaque sous-corpus soit bien contrasté sur le plan thématique, nous avons supprimé toutes les sous-catégories qui relient deux catégories racines ainsi que les articles rattachés à ces catégories (par exemple, les sous-catégories *Organisation sociale, Industrie* et *Entreprise* sur le graphe 1, qui sont à l'intersection des catégories racines *Économie* et *Société*).

Seuls les articles de plus de 300 caractères et d'au-moins une année d'existence ont été conservés, dans la perspective d'éliminer les ébauches d'articles.

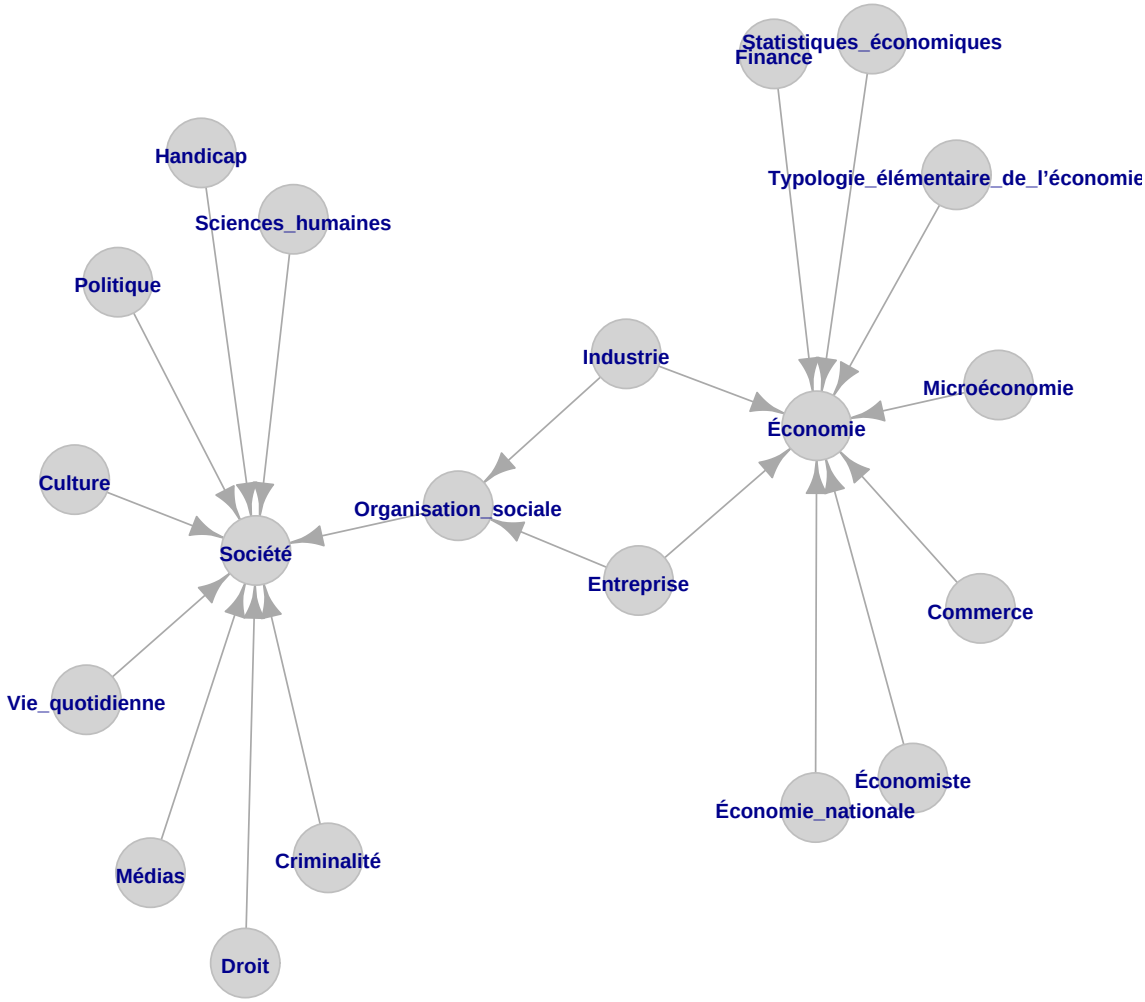

FIG. 1 – Graphe du premier sous-niveau de catégories Wikipédia.

Conversion des articles Nous avons ensuite procédé à une étape de conversion du format des articles, d'abord du format Wikipédia au format TEI[3] par le biais du convertisseur wiki2tei[4], puis du format TEI au format texte brut. La première étape de conversion nous permet de conserver les informations et la syntaxe Wiki tout en bénéficiant d'un format XML plus facilement manipulable. La seconde étape de conversion du format XML au format texte brut a ensuite été réalisée au moyen de requêtes XSLT.

Pour la version finale des articles, nous avons éliminé tous les éléments qui ne relèvent pas directement du contenu textuel des articles (tableaux, listes, bibliographies, tables des matières) ainsi que les titres préformatés utilisés par Wikipédia (« liens externes », « voir aussi », etc).

[3]TEI : Text Encoding Initiative (`www.tei-c.org`), consortium international qui définit et maintient depuis 1994 un standard adapté à la représentation de textes.

[4]Wiki2tei, par Bernard Desgraupes et Sylvain Loiseau : `http://wiki2tei.sourceforge.net/`

3.3 Constitution des corpus

Pour chaque ensemble de catégories relatif à une tâche (voir section 2), nous avons produit les corpus d'apprentissage et de test sur la base d'une répartition respectivement fixée à 60% et 40% du total des articles. Nous avons par ailleurs procédé à une répartition aléatoire des articles en termes de genre et de catégorie d'appartenance dans les différents corpus de test et d'apprentissage.

TÂCHE 1	Appr.	Test
Art	5 767	3 844
Économie	4 630	3 085
Sports	3 474	2 315
Télévision	1 352	1 352

TÂCHE 2	Appr.	Test
France	3 326	2 216
International	5 305	3 536
Littérature	4 576	3 049
Sciences	6 565	4 375
Société	3 778	2 517

FIG. 2 – Proportion d'articles par catégorie pour chacune des tâches.

3.4 Évaluation manuelle de la tâche

Afin de valider le choix de la tâche pour cette édition du défi, nous avons procédé à une évaluation manuelle d'un échantillon du corpus auprès de 4 juges humains. Le corpus à évaluer se composait de 30 articles du *Monde* et 30 articles de Wikipédia. Seuls le titre et le corps de l'article ont été conservés, les tableaux ayant été éliminés de ces fichiers. Les marques distinctives d'appartenance à l'un ou l'autre de ces corpus ont été enlevées (références au *Monde* et bandeaux d'informations utilisés par Wikipédia).

Le test s'est déroulé de la manière suivante : chaque article étant enregistré dans un fichier distinct, les évaluateurs ont eu pour consigne d'identifier le genre et la catégorie sous laquelle chaque article a paru. Tous les articles ont été regroupés en un seul ensemble ; autrement dit, les évaluateurs ont dû choisir parmi toutes les catégories et non parmi deux sous-ensemble de quelques catégories chacun.

Ce test a été réalisé sur une première sélection de 8 catégories :

Le Monde	Wikipédia
Carnet	*Personnalité*
Économie	*Économie*
France	*Politique de la France*
International	*Politique nationale*, moins la catégorie *Politique de la France*
Sciences	*Sciences*
Société	*Société*, moins les sous-catégories *Politique, Personnalité, Sport, Médias*
Sports	*Sport*
Télévision	*Télévision*

FIG. 3 – Correspondance entre catégories du Monde et de Wikipédia sur les 8 catégories utilisées lors du test.

3.4.1 Résultats

Les résultats des juges humains en termes de rappel et précision se sont révélés excellents sur l'identification du genre (F-scores compris entre 0,94 et 1,00) et plutôt bons sur l'identification des catégories (F-scores compris entre 0,66 et 0,82).

	1	**2**	**3**	**4**
Genres	1,00	0,98	0,97	0,94
Catégories	0,79	0,77	0,82	0,66

FIG. 4 – F-scores obtenus par les juges humains sur l'identification du genre et des catégories.

Nous avons par ailleurs confronté ensemble les résultats des juges humains grâce au coefficient κ (Carletta, 1996), coefficient qui permet de mettre en évidence les taux d'accord entre juges.

Juge	**Réf.**	**1**	**2**	**3**	**4**
Réf.		1,00	0,97	0,93	0,87
1	1,00		0,97	0,93	0,87
2	0,97	0,97		0,90	0,83
3	0,93	0,93	0,90		0,87
4	0,87	0,87	0,83	0,87	

Juge	**Réf.**	**1**	**2**	**3**	**4**
Réf.		0,56	0,52	0,60	0,39
1	0,56		0,69	0,75	0,55
2	0,52	0,69		0,71	0,61
3	0,60	0,75	0,71		0,52
4	0,39	0,55	0,61	0,52	

FIG. 5 – Coefficient κ entre juges humains et la référence.
Identification du genre (tableau de gauche) et des catégories (tableau de droite).

Ces coefficients démontrent l'excellent accord des juges entre eux ainsi qu'avec la référence pour l'identification du genre, et font état d'accords modérés à bons pour l'identification des catégories. Ces résultats nous ont confortés dans le choix du thème de ce défi.

3.4.2 Consistance des catégories

Nous avons par ailleurs mesuré la consistance de chaque catégorie en mettant en évidence le rappel et la précision obtenue par l'ensemble des évaluateurs sur chacune des catégories. Cette mesure a été produite sur la base d'une seconde évaluation réalisée par des juges humains, et portant sur un ensemble plus large de catégories (ajout des catégories *Art* et *Littérature*).

La classification des catégories par précisions décroissantes est la suivante : *Sports* (1,00%), *International* (0,80%), *France* (0,76%), *Littérature* (0,76%), *Art* (0,74%), *Télévision* (0,71%), *Économie* (0,58%), *Sciences* (0,33%), *Société* (0,26%). Il en ressort qu'aucun des documents classés dans la catégorie *Sport* n'a été mal classé, alors qu'à l'inverse, les catégories *Sciences* et *Société* sont celles qui ont posé le plus de problèmes.

La classification obtenue sur les rappels décroissants varie légèrement : *International* (0,87%), *Économie* (0,80%), *Sports* (0,75%), *France* (0,70%), *Art* (0,62%), *Littérature* (0,49%), *Télévision* (0,46%), *Société* (0,42%), *Sciences* (0,33%). Ainsi, les articles de la catégorie *International* sont ceux qui ont été le mieux identifiés. Ce classement confirme par ailleurs la difficulté ressentie par les juges humains vis-à-vis des catégories *Société* et *Sciences*.

3.4.3 Ajustement du défi

Les résultats obtenus à l'issue des différents tests réalisés auprès de juges humains nous ont permis, d'une part, de sélectionner les catégories thématiques à conserver pour le défi, et d'autre part, de définir les deux ensembles de catégories utilisés pour chaque sous-tâche.

Alors que nous avions retenu la catégorie *Carnet* (biographies de personnes célèbres) lors du premier test humain, nous avons décidé de l'abandonner dans la suite du défi pour deux raisons. En premier lieu, nous nous sommes rendus compte qu'il s'agissait plutôt d'un genre, le genre « biographie », plutôt qu'une catégorie thématique. D'autre part, nous avons éprouvé quelques difficultés à affecter à une seule catégorie les articles qui pouvaient potentiellement relever de deux catégories, par exemple dans le cas de la biographie d'un sportif qui relèverait des catégories *Carnet* et *Sports*.

En second lieu, nous avons procédé à une répartition des catégories pour chaque sous-tâche sur la base d'un équilibre entre catégories jugées faciles et difficiles par les évaluateurs humains :
– *Art, Économie, Sports, Télévision* pour la sous-tâche combinant reconnaissance de genre et de catégorie ;
– *France, International, Littérature, Sciences, Société* pour la sous-tâche de reconnaissance seule des catégories. Pour ce second ensemble, nous avons fait le choix de rassembler trois catégories assez proches sur le plan thématique (*France, International* et *Société*).

4 Déroulement du défi

Six équipes ont participé à l'édition 2008 du défi, dont une constituée de jeunes chercheurs :
– GREYC (Caen) : F. Rioult, Th. Charnois, Y. Mathet et A. Doucet ;
– LGI2P (Nîmes) et LIRMM (Montpellier) : M. Plantié, G. Dray et M. Roche ;
– LIA (Avignon) : J. M. Torres-Moreno, M. El Bèze, F. Béchet, E. Sanjuan, P. Peinl et P. Bellot ;
– LIA (Avignon) : E. Charton, R. Acuna-Agost, N. Camelin et R. Kessler, *jeunes chercheurs* ;
– LIFO (Orléans) et INaLCO (Paris) : G. Cleuziou et C. Poudat ;
– LIP6 (Paris) : D. Buffoni et A.-P. Trinh.

4.1 Organisation du défi

4.1.1 Corpus d'apprentissage

Les corpus d'apprentissage ont été diffusés à partir du 16 janvier 2008. Comme pour la précédente édition du défi, nous avons autorisé les participants à utiliser des bases de connaissances mais nous avons exclu la possibilité d'utiliser d'autres corpus d'apprentissage que ceux fournis.

4.1.2 Corpus de test

La phase de tests a été élaborée selon les mêmes modalités que l'année précédente : une fenêtre de trois jours comprise dans la période du 17 au 28 mars 2008, les participants ayant le choix du premier jour de cette phase de tests. L'ensemble des candidats s'est porté en faveur de la seconde semaine pour soumettre les résultats.

4.2 Évaluation des résultats

4.2.1 Définition du F-score utilisé pour le classement final

Chaque fichier de résultat a été évalué en calculant le F-score de chacun des corpus avec $\beta = 1$.

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

Lorsque le F-score est utilisé pour évaluer la performance sur chacune des n classes d'une classification, les moyennes globales de la précision et du rappel sur l'ensemble des classes peuvent être évaluées de 2 manières (voir (Nakache & Métais, 2005)) :

– La micro-moyenne qui fait d'abord la somme des éléments du calcul – vrais positifs, faux positifs et négatifs – sur l'ensemble des n classes, pour calculer la précision et le rappel globaux ;
– La macro-moyenne qui calcule d'abord la précision et le rappel sur chaque classe i, puis en fait la moyenne sur les n classes.

Dans la micro-moyenne chaque classe compte proportionnellement au nombre d'éléments qu'elle comporte : une classe importante comptera davantage qu'une petite classe. Dans la macro-moyenne, chaque classe compte à égalité.

Micro-moyenne

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

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.

Les catégories étant inégalement réparties dans les corpus, nous avons choisi de calculer le F-score global avec la macro-moyenne pour que les résultats sur chaque classe comptent de la même manière quelle que soit la taille de la classe.

Par ailleurs, dans la mesure où plusieurs classes peuvent être attribuées au même document avec des indices de confiance, nous avons établi les règles suivantes d'attribution d'une classe à un document pour le calcul du F-score strict.

Un document est attribué à la classe i si :

– Seule la classe i a été attribuée à ce document, sans indice de confiance spécifié ;
– La classe i a été attribuée à ce document avec un meilleur indice de confiance que les autres classes. S'il existe plusieurs classes possédant l'indice de confiance le plus élevé, alors nous retiendrons celle qui sera la première d'entre elles dans la balise `<EVALUATION>`.

Dans le calcul de ce F-score, l'indice de confiance n'est pris en compte que pour sélectionner la catégorie attribuée à un document.

4.2.2 Définition du F-score pondéré par l'indice de confiance

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.

Le F-score pondéré par l'indice de confiance sera utilisé à titre indicatif pour des comparaisons complémentaires entre les méthodes mises en place par les équipes.

Dans le F-score pondéré, 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 attribué 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.

Le F-score pondéré est ensuite calculé à l'aide des formules du F-score classique (voir section 4.2.1).

4.2.3 Algorithme utilisé pour désigner le vainqueur de DEFT'08

Les équipes ont été classées en fonction des rangs obtenus sur l'ensemble des sous-tâches et en considérant chaque soumission comme atomique. Le rang d'une soumission est donc égal à la somme des rangs associés au F-score classique de cette soumission sur chaque sous-tâche. Ainsi, c'est le classement pour chaque sous-tâche qui compte, et non les valeurs cumulées du F-score. L'algorithme utilisé est présenté ci-après :

début
  Pour chaque sous-tâche **faire**
    / Score : liste qui associe à chaque couple (équipe, soumission) son F-score */*
    Score(soumission, équipe) = F-score(sous-tâche, soumission, équipe)
    / Tri de la liste Score dans l'ordre décroissant du F-score */*
    Score trié(soumission, équipe) = tri(Score(soumission, équipe))
    / Tableau des rangs obtenus par chaque soumission de chaque équipe, pour la sous-tâche considérée */*
    Rang[sous-tâche][soumission][équipe] = rang(Score trié(soumission, équipe))
  fin Pour
  Pour chaque équipe ayant soumis **faire**
    / Somme, sur toutes les sous-tâches, des rangs obtenus pour chaque soumission */*
    Rang global[soumission][équipe] = $\sum_{\text{sous-tâche}}$ rangs[sous-tâche][soumission][équipe]
    / Choix de la meilleure soumission (rang le plus faible) */*
    Rang[équipe] = $\min_{\text{soumission}}$(rangs[soumission][équipe])
  fin Pour
  / Choix du vainqueur : équipe dont le rang est le plus faible */*
  ÉquipeV telle que : Rang[ÉquipeV] = $\min_{\text{équipe}}$(Rang[équipe])
fin

FIG. 6 – Algorithme pour désigner le vainqueur

5 Conclusion

Dans cet article, nous avons présenté les différents corpus utilisés et les méthodes que nous avons mobilisées pour constituer les corpus de test et d'apprentissage. Nous avons également détaillé les différents tests qui ont été réalisés entre juges humains, ces tests nous ayant permis, d'une part d'affiner la tâche de cette campagne, et d'autre part de mesurer la faisabilité de la tâche. Enfin, nous avons rappelé les différentes étapes du déroulement de ce défi en insistant notamment sur les mesures utilisées pour évaluer et classer les résultats des participants.

Remerciements

Nous remercions la société ELDA pour la mise à disposition grâcieuse du corpus du Monde et nos partenaires : le CNRS, l'AFIA, l'ATALA et Wikipédia.

Références

CARLETTA J. (1996). Assessing agreement on classification tasks : the kappa statistics. *Computational Linguistics*, **2**(22), 249–254.

NAKACHE D. & MÉTAIS E. (2005). Evaluation : nouvelle approche avec juges. In *INFORSID*, p. 555–570, Grenoble.