Présentation de DEFT'07 (DÉfi Fouille de Textes)

Les membres du Comité d'Organisation de DEFT'07 :
Cyril Grouin[1], Jean-Baptiste Berthelin[1], Sarra El Ayari[1], Thomas Heitz[2], Martine Hurault-Plantet[1], Michèle Jardino[1], Zohra Khalis[3] et Michel Lastes[1]

[1] LIMSI–CNRS,
{cyril.grouin, jean-baptiste.berthelin, sarra.elayari, martine.hurault-plantet, michel.lastes}@limsi.fr

[2] LRI, Université Paris-Sud,
heitz@lri.fr

[3] Épigénomique, Génopole d'Évry,
zkhalis@epigenomique.genopole.fr

Résumé : Le thème de cette édition du Défi Fouille de Textes est la classification de textes d'opinion. Pour réaliser ce défi, nous avons rassemblé quatre corpus venant de domaines différents, critiques de spectacles, tests de jeux, relectures d'articles scientifiques et débats sur des projets de loi. Dans cet article, nous présentons les corpus, ainsi que les pré-traitements de nettoyage que nous avons dû effectuer. Nous décrivons ensuite les tests manuels de la tâche de classification en valeurs d'opinion que avons effectués dans le but d'évaluer sa faisabilité sur nos corpus. Nous décrivons enfin les scores utilisés pour l'évaluation des résultats.

1 Introduction

Après le succès des éditions précédentes du défi fouille de texte (DEFT) consacrées à l'identification de locuteurs dans des discours politiques en 2005, et à la segmentation thématique de textes politiques, scientifiques et juridiques en 2006 (voir Azé *et al.* (2006)), une troisième édition a été programmée pour l'année 2007. Cette édition s'inscrit dans le cadre de la plate-forme de l'Afia (Association Française d'Intelligence Artificielle) organisée à Grenoble du 2 au 6 juillet 2007.

Le thème retenu cette année concerne l'attribution automatique de valeurs d'opinion à des textes présentant un avis argumenté, positif ou négatif, sur un sujet donné.

2 Présentation des corpus

Corpus « À voir, à lire »

Ce corpus comprend environ 3 000 documents (7,6 Mo), pour l'essentiel des critiques de livres, complétés par des critiques de films et de spectacles. Ces documents proviennent du site Internet www.avoir-alire.com. Trois valeurs d'opinion sont proposées pour ce corpus : favorable (classe 2), neutre (classe 1) et défavorable (classe 0).

Critiques de jeux vidéos

Ce corpus se compose d'environ 4 000 critiques de jeux vidéos (28,3 Mo) portant sur divers aspects du jeu (graphisme, jouabilité, durée, son, etc.) et provenant du site Internet www.jeuxvideos.com. Trois valeurs d'opinion sont proposées pour ce corpus : appréciation positive du jeu vidéo (classe 2), appréciation moyenne (classe 1) et appréciation négative (classe 0).

Relectures d'articles scientifiques

Ce corpus intègre environ 1 000 relectures d'articles (2,4 Mo) relatifs au domaine de l'Intelligence Artificielle. Ces relectures sont issues des conférences JADT[1], RFIA[2] et TALN[3]. Trois valeurs d'opinion sont proposées pour ce corpus : article accepté en l'état ou après modifications mineures (classe 2), article accepté après modifications majeures (classe 1) et article rejeté (classe 0).

Débats parlementaires

Ce corpus regroupe 28 832 interventions de Députés à l'Assemblée Nationale (38,1 Mo) extraites des débats portant sur la loi relative à l'énergie. Ces débats ont été aspirés depuis le site Internet de l'Assemblée Nationale[4]. Contrairement aux précédents corpus, seules deux valeurs d'opinion sont disponibles pour ce corpus : vote favorable à la loi en examen (classe 1) et vote défavorable à la loi en examen (classe 0).

Seuls les textes ont été retenus dans la composition des documents de chaque corpus, toute autre information (images, tableaux, etc.) ayant été supprimée.

Chaque corpus a été segmenté en corpus d'apprentissage et de test sur la base d'une répartition à 60 et 40%.

3 Préparation des données

3.1 Traitements spécifiques effectués sur chaque corpus

De manière générale et pour chaque corpus, des phases de nettoyage se sont révélées nécessaires (encodage des caractères en ISO Latin-1, éliminations des accents sur les caractères en dehors de l'ISO Latin-1 : « ō » devenant ainsi « o » dans « Tōkyō », homogénéisation des fins de ligne) puis de conversion des documents au format XML. Une DTD a été réalisée pour l'ensemble des corpus du défi.

Corpus « À voir, à lire » et « Jeux vidéos »

Les traitements ont été globalement les mêmes pour ces deux corpus. Tout d'abord il a fallu aspirer[5] les pages des critiques. Les pages HTML obtenues ont ensuite été analysées pour en extraire uniquement les textes des critiques et les notes qui leur sont associées.

Relectures

Les documents d'origine de ce corpus ont été rédigés dans des traitements de texte ou des tableurs. Ils ont été convertis au format texte brut avec conversion de l'encodage et élimination des formules mathématiques LaTeX. L'ensemble des documents a été anonymisé. Seules les relectures rédigées en français ont été conservées dans le corpus des JADT[6].

Débats parlementaires

Les compte-rendus des débats parlementaires ont été aspirés depuis le site Internet de l'Assemblée Nationale. Les questions au Gouvernement ont été manuellement retirées de ces comptes-rendus, qui ont ensuite été formatés en XML. Effectuer la correspondance entre l'intervention d'un locuteur et la valeur du vote de ce locuteur n'a pas posé de problème particulier étant donné que chaque compte-rendu de l'Assemblée Nationale reprend, en préambule, la liste des votes des parlementaires.

Pour éviter tout biais, le contenu des interventions a été anonymisé sur la base des noms de personnes (250 hommes politiques), de lieux (une dizaine de métonymies politiques : *l'Élysée, Matignon, Place Beauvau, rue de Grenelle*) et de partis politiques (*UMP, PS, droite, gauche, républicain, extrême droite*).

Seules les interventions de plus de 300 caractères ont été conservées pour le défi, les document en-deçà de ce seuil n'ayant pas été jugés exploitables après les tests réalisés auprès de juges humains.

[1] Journées internationales d'Analyse statistique des Données Textuelles.

[2] Reconnaissance des Formes et Intelligence Artificielle.

[3] Traitement Automatique des Langues Naturelles.

[4] L'intégralité des séances de débats sur ce projet de loi est accessible à l'adresse `http://www.assemblee-nationale.fr/12/debats/`

[5] Après accord des propriétaires des sites, bien évidemment.

[6] Ce corpus comprenant des relectures rédigées en anglais, en français ou en italien.

3.2 Principales difficultées rencontrées

Du fait de l'hétérogénéité des sources des différents documents composant nos corpus, nous avons dû faire face à plusieurs formats de documents (pages web, documents Word, tableaux Excel, fichiers en texte brut). Le premier « défi » qui s'est imposé a été celui de la conversion de l'ensemble de ces documents en fichiers exploitables pour la suite de la campagne. Il ne nous a pas toujours été donné de convertir automatiquement les documents, en particulier dans le cas des tableaux réalisés sous Excel.

Un second problème, fortement lié au point précédent, concerne les encodages de caractères et des fins de ligne. Les documents rédigés au moyen du traitement de textes Word intègrent notamment quelques caractères encodés en UTF-8 tels que : le symbole de l'euro (codé 200 en octal), les points de suspension (codés 205 en octal), la ligature « œ » (codée 234 en octal pour la version en minuscules et 214 pour la version en majuscules) et les guillemets simples « à l'anglaise » ouvrantes et fermantes (codées 221 et 222 en octal).

Malgré ces précautions, il reste probablement quelques coquilles dans nos corpus. Mais, après tout, cela fait partie des difficultés du traitement de la langue naturelle.

3.3 Évaluations manuelles des corpus

Chacun des corpus proposés dans le cadre de ce défi a auparavant été testé auprès de juges humains qui ont eu pour charge d'attribuer une valeur à quelques extraits des quatre corpus. Les résultats de chacun des juges ont été confrontés par le biais du coefficient κ (Kappa) de Cohen (1960) qui permet de mettre en évidence le taux d'accord entre deux juges[7].

Juge	**Réf.**	**1**	**2**
Réf.		0,17	0,12
1	0,17		0,03
2	0,12	0,03	

Juge	**Réf.**	**1**	**2**
Réf.		0,74	0,79
1	0,74		0,74
2	0,79	0,74	

FIG. 1 – Coefficient κ entre juges humains et la référence sur le corpus des jeux vidéos.
Échelle de notes de 0 à 20 (tableau de gauche) et de 0 à 2 (tableau de droite).

Les évaluations humaines ont permis de tester différentes échelles de notes. Les tableaux n° 1 donnent ainsi les coefficients κ obtenus par deux juges humains – entre eux et vis-à-vis de la référence – pour le corpus des jeux vidéos selon deux échelles de notes : une échelle large de 0 à 20 (notes d'origine) pour le tableau de gauche et une échelle restreinte de 0 à 2 pour le tableau de droite. Le changement d'échelle est le suivant : classe 0 de 0 à 9, classe 1 de 10 à 14 et classe 2 de 15 à 20.

Ces résultats démontrent qu'il y a un mauvais accord entre les juges sur l'échelle large (coefficient κ inférieur à 0,20) tandis que l'accord est qualifié de « bon » sur l'échelle restreinte (coefficient κ compris entre 0,61 et 0,80). Le mauvais accord entre juges sur l'échelle large s'explique par la dispersion des notes de 0 à 20.

Juge	**Réf.**	**1**	**2**	**3**	**4**	**5**
Réf.		0,10	0,29	0,39	0,46	0,47
1	0,10		0,37	0,49	0,48	0,35
2	0,29	0,37		0,36	0,30	0,43
3	0,39	0,49	0,36		0,49	0,54
4	0,46	0,48	0,30	0,49		0,60
5	0,47	0,35	0,43	0,54	0,60	

Juge	**Réf.**	**1**	**2**	**3**	**4**	**5**
Réf.		0,27	0,62	0,53	0,56	0,67
1	0,27		0,45	0,43	0,57	0,37
2	0,62	0,45		0,73	0,48	0,54
3	0,53	0,43	0,73		0,62	0,62
4	0,56	0,57	0,48	0,62		0,76
5	0,67	0,37	0,54	0,62	0,76	

FIG. 2 – Coefficient κ entre juges humains et la référence sur le corpus « *à voir, à lire* ».
Échelle de notes de 0 à 4 (tableau de gauche) et de 0 à 2 (tableau de droite).

Ces différences d'accord entre juges se retrouvent sur l'ensemble des corpus composant cette édition du défi. Les tableaux n° 2 renseignent des coefficients κ obtenus par cinq juges sur le corpus « *à voir, à lire* », pour deux échelles de notes : une échelle large (de 0 à 4) pour le tableau de gauche et une échelle restreinte (de 0 à 2) pour le

[7] L'accord entre deux juges est ainsi qualifié selon la valeur prise par le coefficient κ : excellent de 0,81 à 1,00 – bon de 0,61 à 0,80 – modéré de 0,41 à 0,60 – médiocre de 0,21 à 0,40 – mauvais de 0 à 0,20 – très mauvais en négatif.

tableau de droite. Le changement d'échelle est le suivant : classe 0 de 0 à 1, classe 1 pour la note 2 et classe 2 pour les notes 3 à 4.

À l'instar du corpus des jeux vidéos, les accords entre juges sur le corpus « *à voir, à lire* » sont meilleurs sur une échelle restreinte (de 0 à 2) que sur l'échelle large (de 0 à 4). Sur l'échelle large, les coefficients κ sont compris entre 0,10 et 0,60 (accords mauvais à modérés) tandis qu'ils s'échelonnent entre 0,27 et 0,76 (accords médiocres à bons) sur l'échelle restreinte.

Suite à ces évaluations manuelles, nous avons choisi d'utiliser des échelles restreintes pour l'ensemble des corpus du défi : une échelle de 0 à 2 pour les corpus « *à voir, à lire* », des jeux vidéos et des relectures, et une échelle de 0 à 1 pour le corpus des débats parlementaires (voir tableau n° 3).

	A voir, à lire	**Jeux vidéos**	**Relectures**	**Débats**
0	Mauvais	Mauvais	Article rejeté	Contre la loi
1	Moyen	Moyen	Article accepté après modifications majeures	Pour la loi
2	Bon	Bon	Article accepté en l'état ou après modifications mineures	

FIG. 3 – Valeurs associées à chaque classe selon les corpus.

D'autre part, les évaluateurs humains ont jugé la tâche plus facile sur les corpus des jeux vidéo et des débats parlementaires que pour le corpus « *à voir, à lire* ». Les coefficients κ sur les échelles restreintes sont également meilleurs pour les deux premiers corpus que pour le dernier ; ils sont compris entre 0,74 et 0,79 pour le corpus des jeux vidéos (bon accord), entre 0,60 et 0,80 pour le corpus des débats parlementaires (bon accord) et entre 0,27 et 0,76 pour le corpus « *à voir, à lire* » (donc, des accords médiocres à modérés).

3.4 Indice de confiance

Définition

Un système 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 classe donnée.

Usage

L'indice de confiance introduit une pondération de la confiance et du rappel, donc du F-score. Il permet donc de comparer des classifieurs sur une base différente du « tout ou rien ».

Pertinence dans DEFT'07

Nous avons proposé aux concurrents un recours optionnel à cette variable.

Si l'on examine la situation dans les quatre corpus, on constate que son utilité n'est sans doute pas la même dans chacun des cas. Les critiques de films, par exemple, sont souvent identifiables assez nettement comme soit favorables, soit défavorables. Au contraire, dans le cas des relectures d'articles scientifiques, les documents sont plus difficiles à interpréter, et il semble alors légitime de pouvoir représenter le fait qu'un même jugement contient « du pour » et « du contre ».

De même, dans les débats parlementaires, certaines interventions ne se laissent pas facilement catégoriser comme purement favorables ou défavorables à un projet de loi, les attitudes mitigées sont possibles.

Les valeurs de l'indice de confiance peuvent traduire cette difficulté relative à choisir entre deux ou trois classes. Si elles avoisinent 0 ou 1, on est dans un cas tranché, si elles sont de l'ordre d'un demi ou d'un tiers, il s'agit d'un cas plutôt équivoque.

Parmi les concurrents, certains ont, dans un premier temps, mis en compétition plusieurs classifieurs dont ils disposaient. Ceux qui produisaient des jugements bien tranchés étaient préférés à ceux qui se montraient perplexes. C'est précisément grâce à leurs indices de confiance qu'une telle comparaison a pu s'effectuer.

4 Déroulement du défi

Les équipes ayant participé au défi sont au nombre de dix dont trois équipes constituées uniquement de jeunes chercheurs :

- **CELI France (Grenoble) :** Sigrid Maurel, Paolo Curtoni et Luca Dini ;
- **EPHE (Paris) et Universität Würzburg (Würzburg, Allemagne) :** Murat Ahat, Wolfgang Lenhard, Herbert Baier, Vigile Hoareau, Sandra Jhean-Larose et Guy Denhière ;
- **GREYC (Caen) :** Matthieu Vernier, Yann Mathet, François Rioult, Thierry Charnois, Stéphane Ferrari et Dominique Legallois ;
- **Lattice (Paris) :** Alejandro Acosta et André Bittar, *équipe jeunes chercheurs* ;
- **LGI2P (Nîmes) et LIRMM (Montpellier) :** Michel Plantié, Gérard Dray et Mathieu Roche ;
- **LIA (Avignon) :** Juan Manuel Torres-Moreno, Marc El-Bèze, Frédéric Béchet et Nathalie Camelin ;
- **LIA (Avignon) :** Éric Charton et Rodrigo Acuna-Agost, *équipe jeunes chercheurs* ;
- **LIP6 (Paris) :** Anh-Phuc Trinh, *équipe jeune chercheur* ;
- **NLTG–Université de Brighton (Royaume-Uni) :** Michel Généreux et Marina Santini ;
- **Yahoo ! Inc. (Paris) :** Eric Crestan, Stéphane Gigandet et Romain Vinot.

4.1 Organisation du défi

Corpus d'apprentissage

Les corpus d'apprentissage ont été diffusés à partir du 4 janvier 2007. Il a été autorisé aux différents participants d'utiliser des bases de connaissances. En revanche, nous avons exclu la possibilité d'utiliser des corpus d'apprentissages autres que ceux que nous avons fournis.

Nous donnons ci-dessous un extrait du corpus d'apprentissage des débats parlementaires (les passages anonymisés de ce corpus ont été remplacés par des balises) :

```
<DOCUMENT id="4:6">
 <EVALUATION nombre="1">
  <NOTE valeur="0" confiance="1.00" />
 </EVALUATION>
 <TEXTE>
 <![CDATA[
 Au nom de cette nouvelle gouvernance, vous affirmez la nécessité d'un grand
 nombre de réformes dans l'Etat, et notamment d'une nouvelle étape de la dé-
 centralisation. Les <partiPolitique /> qui, en 1982, avec <hommePolitique />
 et <hommePolitique />, ont élaboré et voté les grandes lois de décentrali-
 sation contre une <partiPolitique /> qui y voyait une menace contre l'unité
 de la République et un affaiblissement de l'Etat, ne peuvent que partager
 cette perspective. Sur ce socle, vous proposez d'organiser notre administra-
 tion selon un nouveau schéma. Pour une part, il s'agit de constitutionnali-
 ser une institution comme la région – qui pourrait sérieusement s'y opposer ?
 – de reconnaître le principe de l'autonomie financière des collectivités lo-
 cales, et d'introduire les référendums locaux : autant de thèmes sur lesquels
 nous pouvons converger. Enfin vous voulez faire droit au principe d'expéri-
 mentation. Nous y avons nous-mêmes recouru.
 ]]>
 </TEXTE>
</DOCUMENT>
```

Quatre équipes se sont désistées, trois avant la phase de tests, l'une pendant la phase de tests, ce qui constitue un taux d'abandon de 28,6%.

Corpus de test

La phase de tests a été conçue sous la forme d'une fenêtre de trois jours à définir dans un délai de deux semaines – du 19 au 30 mars 2007 –, les candidats ayant dès lors toute latitude pour choisir le premier jour du test dans cette période.

Les onze équipes ayant participé au test ont toutes choisi la deuxième semaine pour soumettre leurs 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 $\beta = 1$.

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

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

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

Macro-moyenne

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

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 classes d'opinion 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 classe d'opinion 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 classe d'opinion 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'07

Les équipes ont été classées en fonction des rangs obtenus sur l'ensemble des corpus 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 corpus. Ainsi, c'est le classement pour chaque corpus qui compte, et non les valeurs cumulées du F-score.

L'algorithme utilisé est présenté ci-dessous :

début
Pour chaque corpus (corpus ∈ {à voir à lire, jeux, relectures, débats}) **faire**
/ Score : liste qui associe à chaque couple (équipe, soumission) son F-score */*
Score(soumission, équipe) = F-score(corpus, 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 le corpus considéré */*
Rang[corpus][soumission][équipe] = rang(Score trié(soumission, équipe))
fin Pour
Pour chaque équipe ayant soumis **faire**
/ Somme, sur tous les corpus, des rangs obtenus pour chaque soumission */*
Rang global[soumission][équipe] = $\sum_{\text{corpus}}$ rangs[corpus][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. 4 – Algorithme pour désigner le vainqueur

5 Conclusion

Cet article présente l'édition 2007 du défi fouille de textes dont l'objectif vise à attribuer automatiquement une classe à un texte d'opinion relevant de trois thématiques différentes (média, scientifique et juridique). L'ensemble des documents composant nos corpus est rédigé en français.

Après avoir exposé la tâche à réaliser et présenté les corpus utilisés pour ce défi, nous avons décrit les étapes de préparation des corpus en mettant l'accent sur les traitements spécifiques effectués sur chaque corpus et sur les problèmes rencontrés. Nous avons ensuite détaillé le déroulement du défi en présentant notamment la procédure d'évaluation des résultats et de classement des équipes.

Références

AZÉ J., HEITZ T., MELA A., MEZAOUR A.-D., PEINL P. & ROCHE M. (2006). Présentation de DEFT'06 (DÉfi Fouille de Textes). In *Actes de DEFT'06*.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, (20), 37–46.

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