DEFT'07 : une campagne d'évaluation en fouille d'opinion

Cyril Grouin*, Martine Hurault-Plantet*
Patrick Paroubek*, Jean-Baptiste Berthelin*

*LIMSI–CNRS
BP133 – F-91403 Orsay Cedex
{Cyril.Grouin, Martine.Hurault-Plantet, Patrick.Paroubek, Jean-Baptiste.Berthelin}@limsi.fr,
http://deft.limsi.fr/

Résumé. Depuis 2005, les campagnes nationales d'évaluation « DEFT » proposent des thématiques de recherche exploratoires axées sur la fouille de texte. L'édition 2007 a porté sur la classification de textes d'opinion : la tâche consistait à attribuer une classe d'opinion à chaque texte d'un corpus, parmi 2 ou 3 classes allant d'un jugement défavorable à un jugement favorable. Quatre corpus ont été mis à la disposition des participants : débats parlementaires sur un projet de loi, critiques de jeux vidéos, critiques de films et de livres, et relectures d'articles de conférences. Dans cet article, nous décrivons d'abord la phase préparatoire de la campagne, avec la collecte des corpus, la définition des mesures d'évaluation, et des tests humains de la tâche. Nous présentons ensuite une analyse des résultats des participants, et les remarques qui en découlent concernant les différents types de corpus. Enfin, nous faisons un bilan synthétique des méthodes proposées à l'évaluation.

Introduction

Dans le domaine du traitement automatique du langage, s'il existe depuis longtemps des campagnes d'évaluation anglophones récurrentes, c'est chose beaucoup plus rare en France. Celles qui ont eu lieu jusqu'ici, telles que GRACE (voir Adda et al., 1998) ou l'ensemble des campagnes TECHNOLANGUE (Mapelli et al., 2004), ont été organisées sous la forme de projets soumis à des appels d'offre, par essence limités dans le temps. Une exception notoire est constituée, au niveau de l'Europe, par la campagne d'évaluation CLEF (Peters, 2000; Braschler et Peters, 2004), qui est multilingue, subventionnée par la CE depuis sa première édition en 2000, et qui, à l'instar de TREC[1], gère plusieurs tâches.

Le défi fouille de textes (DEFT) a été créé en 2005 par un groupe de chercheurs (Prince et al., 2007), dans le but d'initier une série de campagnes d'évaluation francophones sur des thématiques relevant de la fouille de textes. Organisé d'abord par le LRI[2] puis par le LIMSI[3], ce défi est proposé tous les ans sur une thématique différente. Chaque édition a rassemblé

[1] `http://trec.nist.gov`
[2] Laboratoire de Recherche en Informatique, `http://www.lri.fr`
[3] Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur `http://www.limsi.fr`

une dizaine d'équipes participantes, pour la plupart issues des laboratoires de recherche publics français. Les campagnes DEFT visent essentiellement à mettre à disposition à la fois des protocoles et des outils de mesure des systèmes d'analyse du langage, et des corpus d'expérimentation en français, sur le long terme.

Une campagne d'évaluation se construit en plusieurs étapes dont la première est la définition du thème de la campagne et de la tâche à réaliser par les participants. Les mesures permettant l'évaluation des méthodes et logiciels réalisant la tâche sont ensuite examinées. L'étape suivante concerne la collecte des corpus ainsi que des données de référence de la tâche (Adda et al., 1999). Les données de référence sont les résultats attendus de la tâche auxquels seront comparés les résultats des participants. Ces données de référence sont souvent établies par des juges humains, soit a priori, par une annotation des corpus avant la campagne d'évaluation, soit a posteriori, par un jugement des résultats des participants. Mais elles peuvent aussi se trouver d'emblée dans le corpus ou dans ses méta-données, et c'est le cas par exemple des critiques de films qui contiennent à la fois un texte évaluatif constitué par la critique elle-même, et une note globale sous des formes iconiques variées, étoiles, smileys ou autre, qui résume l'opinion exprimée dans le texte. Par nécessité, c'est cette deuxième solution, choisir des corpus qui contiennent les données de référence de la tâche, qui a été adoptée dans les campagnes DEFT. Néanmoins, la préparation de la campagne comporte un test de la tâche par des juges humains, sur un petit échantillon du corpus. Les résultats de ce test nous permettent de mesurer la difficulté de la tâche, et éventuellement de faire évoluer sa définition. Après la campagne proprement dite, un atelier est consacré au bilan des résultats, permettant une comparaison des différentes méthodes mises en œuvre par les participants, et l'étude de leurs avantages et inconvénients respectifs (Grouin et al., 2007).

Dans cet article, nous nous proposons de présenter les différents aspects de la campagne d'évaluation DEFT sur la fouille d'opinion. Dans un premier temps, nous décrirons les étapes liées à la préparation de la campagne, du point de vue de la constitution des corpus, et des problèmes qui se sont présentés ainsi que les solutions adoptées. Nous détaillerons ensuite les évaluations manuelles réalisées par le Comité d'organisation, puis nous présenterons les mesures d'évaluation retenues pour la campagne. Enfin, nous ferons une analyse des résultats de la campagne ainsi qu'un bilan des méthodes utilisées par les participants de ce défi.

1 La fouille d'opinion et son évaluation

Les sondages d'opinion, fondés sur des enquêtes statistiques très cadrées, existent depuis longtemps. En revanche, la fouille d'opinion, beaucoup plus récente, est née d'Internet, des journaux en ligne, et du Web 2.0. Ces deux domaines se rejoignent dans leurs buts et dans leurs applications, essentiellement prises dans le marketing et la politique. Ils diffèrent néanmoins par leurs méthodologies de collecte des opinions. Les sondages reposent sur des opinions émises en réponse aux questions d'une enquête sur un sujet bien délimité, par des personnes appartenant à un échantillon préalablement construit. La fouille d'opinion prend en général ses informations dans des opinions émises spontanément. Mais dans les deux cas, il faut être capable de mesurer l'opinion exprimée au regard d'une échelle de valeurs définie à l'avance.

Les gisements d'information sous format électronique semblant presque illimités, les corpus potentiels qu'ils représentent font naître un intérêt croissant pour les méthodes qui per-

mettent de les utiliser. La fouille d'opinion représente donc un domaine de recherche très actuel autant par ses enjeux applicatifs que par les méthodes développées (Pang et Lee, 2008), et elle a suscité plusieurs campagnes d'évaluations.

1.1 Les campagnes d'évaluation en fouille d'opinion

Plusieurs campagnes de fouille d'opinion ont eu lieu à partir de 2006. La campagne TREC[4] (Text Retrieval Conference) a organisé entre 2006 et 2008 une tâche de recherche d'information dans la blogoshère, la Blog Track[5], qui comportait une tâche de recherche d'opinion. Cette tâche avait deux objectifs, une séparation des posts de blogs en objectif/subjectif (ceux qui expriment, ou non, une opinion sur une cible donnée), et une séparation des posts en opinion positive/négative, avec un classement dans l'ordre de positivité/négativité décroissante.

La campagne NTCIR-MOAT[6] (Multilingual Opinion Analysis Task) a lieu tous les ans depuis 2006. Les principales langues concernées sont le japonais, le chinois et l'anglais. La tâche consiste en un étiquetage très détaillé d'articles de journaux. Chaque phrase doit être étiquetée comme étant soit objective soit subjective, et pertinente ou non par rapport à un thème donné. Le porteur et la cible de l'opinion doivent également être étiquetés.

En 2007, SEMEVAL[7] a proposé une tâche *Affective text* qui consistait à catégoriser des titres de journaux d'une part suivant l'émotion exprimée, parmi six émotions de base (colère, dégoût, peur, joie, tristesse, surprise), et d'autre part suivant la valence positive ou négative de l'émotion exprimée.

Enfin, en 2008, la campagne TAC[8] (Text Analysis Conference) a comporté une tâche de type question-réponse avec des questions d'opinion sur un thème donné. Un premier type de question portait sur les porteurs d'opinion (*qui soutient quoi ?*), et un deuxième type de questions portait sur l'opinion elle-même (*quelles sont les critiques sur ..., pourquoi les gens aiment ...*). La tâche comportait aussi un résumé automatique d'opinion qui devait être élaboré à partir des réponses à la tâche question-réponse d'opinion.

Les performances des participants à ces campagnes ont été assez différents suivant la tâche. La détermination de la polarité n'a pas dépassé une F-mesure de 0,51 (voir la section 4.1 pour la définition de la F-mesure) et la reconnaissance du porteur de l'opinion a été en général moins performante. C'est la tâche de catégorisation des titres de journaux suivant l'émotion exprimée qui a obtenu les plus faibles résultats, avec une F-mesure comprise entre 0,02 et 0,30 pour la tristesse (émotion la mieux reconnue), et une F-mesure nulle pour le dégoût (émotion la moins bien reconnue).

1.2 La campagne DEFT

Détermination de la tâche Le thème de la fouille d'opinion est vaste. En effet, une analyse d'opinion commence par la détection du caractère plus ou moins subjectif d'un texte ou d'un passage, c'est-à-dire par déterminer s'il est porteur d'un « sentiment », d'un jugement, d'une opinion, ou au contraire de données essentiellement factuelles. Les parties de texte qui

[4] http://trec.nist.gov
[5] http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG/
[6] http://ntcir.nii.ac.jp/index.php/Table/MOAT/
[7] http://nlp.cs.swarthmore.edu/semeval/tasks/task14/summary.shtml
[8] http://www.nist.gov/tac/tracks/2008/qa/

contiennent une opinion sont ensuite analysées pour donner une valeur à l'opinion exprimée, soit suivant une polarité positive/négative, soit suivant une échelle de valeurs. Enfin, le jugement exprimé sur un sujet particulier peut être influencé par, ou laisser transparaître, des opinions d'un type plus général comme par exemple une opinion politique. Nous avons donc là tout un éventail de tâches possibles. Le thème de la campagne 2007 de DEFT a porté sur le problème de la mesure de l'opinion exprimée suivant une échelle de valeurs finies de 2 à 3 niveaux allant d'un jugement favorable à un jugement défavorable. Étant donné un corpus de textes d'opinion, la tâche consistait donc à attribuer une classe d'opinion à chaque texte de ce corpus.

En tant qu'objet de recherche, la fouille d'opinion fait appel à des méthodes statistiques aussi bien que linguistiques. En effet, pour traiter de grandes quantités de textes, repérer des régularités, structurer par domaine, thème, opinion, les méthodes statistiques sont efficaces. Mais les méthodes linguistiques s'avèrent performantes pour modéliser des comportements langagiers particuliers et les détecter ensuite dans les textes. Elles ont également l'avantage d'être explicatives du phénomène mis en jeu. L'enjeu de la campagne était donc d'avoir des éléments de comparaison de l'application de ces différentes méthodes.

Déroulement de la campagne Pour les participants, la campagne d'évaluation s'est déroulée en deux temps. Les équipes participantes ont d'abord pu affiner leurs méthodes sur des corpus d'apprentissage, pendant une période d'environ trois mois. Elles ont ensuite eu trois jours pour l'évaluation proprement dite de leurs logiciels sur les corpus de test.

Les équipes ont été autorisées à utiliser des bases de connaissances. En revanche, elles ne devaient utiliser comme corpus d'apprentissage que ceux fournis pour la campagne.

Finalement, nous avons eu dix équipes participantes qui ont appliqué leurs logiciels sur les corpus de test. Parmi ces dix équipes, huit venaient de laboratoires universitaires. Les deux autres venaient d'entreprises privées, l'une spécialisée dans la fouille d'opinion et l'autre dans la recherche d'information. Trois des équipes universitaires étaient constituées uniquement de jeunes chercheurs (doctorants ou détenteurs d'une thèse depuis moins d'un an).

2 Collecte des corpus

Une évaluation n'est pas un absolu. Elle est toujours relative aux corpus choisis et aux mesures d'évaluation considérées. Pour cette raison, il est préférable de proposer plusieurs corpus différents pour une même tâche, et les mesures d'évaluation que nous avons considérées sont celles communément admises dans la communauté visée par le thème de la campagne.

Nous avons donc cherché à rassembler des corpus de sources différentes, répondant à deux critères principaux : en premier lieu, la capacité d'accès aux données – tant du point de vue de la récupération que de celui de la redistribution des données – et en second lieu, la possibilité d'extraire du corpus les données de référence de la tâche. Nous avons finalement obtenu quatre corpus : un corpus de critiques de films et de livres et un corpus de critiques de jeux vidéos comportant pour chaque critique à la fois un texte évaluatif et une note globale appréciative, un corpus de relectures d'articles de conférences comportant le texte évaluatif et la notification d'acceptation ou de rejet, et enfin un corpus de débats parlementaires sur des projets de lois auquel nous avons pu associer à chaque intervenant dans les débats les méta-données de son vote pour ou contre le projet de loi. Ces corpus sont décrits plus en détail dans la section 2.3.

2.1 Accès aux données

Problèmes juridiques L'utilisation de corpus dans une campagne d'évaluation soulève des problèmes juridiques de redistribution des données qui se posent différemment suivant les différents statuts de ces données.

Nous avons constitué un premier corpus, composé de débats parlementaires qui se sont tenus à l'Assemblée Nationale sur un projet de loi relative à l'énergie. L'Assemblée Nationale étant une institution publique, la redistribution des données présentes sur son site Internet s'avère possible sans condition. Les débats parlementaires s'accompagnant d'un vote individuel, la valeur de ces votes (pour ou contre le projet de loi) nous a permis de réaliser la référence.

Un second type de corpus a été produit en rassemblant des critiques de livres, de films et de jeux vidéos provenant de sites Internet spécialisés[9]. Pour ces corpus, nous avons sollicité les responsables des sites pour pouvoir en aspirer le contenu et le redistribuer dans le cadre de la campagne DEFT. Les sites que nous avons retenus présentaient l'avantage de gérer à la fois l'hébergement du contenu et sa production, donc d'en être complètement propriétaires. En effet, certains sites gèrent l'hébergement mais sans gérer la production du contenu, qui est alors fourni soit par des internautes, soit par des liens vers des sites eux-mêmes producteurs de contenu. Dans ce cas, se pose le problème juridique de la réutilisation de données hébergées par le site mais déposées par des tiers (tels que les sites participatifs où les consommateurs donnent leur avis sur différents types de produits).

Enfin, nous avons également produit un corpus sur la base de relectures d'articles scientifiques soumis à des conférences nationales françaises du domaine du traitement automatique des langues (TAL). Ce type de corpus pose cependant la question de la sensibilité des données, les articles critiqués pouvant éventuellement être reconnus. L'utilisation et la redistribution de ces données n'a pu être possible qu'après une phase importante d'anonymisation de toute référence permettant de remonter à l'auteur de l'article, ou à son relecteur.

Les problèmes juridiques liés à la redistribution des données constituent un problème important dans la préparation des campagnes d'évaluation, car il n'existe finalement qu'assez peu de données récupérables et redistribuables sans contraintes juridiques. Très récemment, les organisateurs de la campagne d'évaluation MOAT (conférence NTCIR-7) ont dû renoncer temporairement, pour des raisons juridiques, à constituer un corpus de blogs[10].

Récupération des données Les campagnes DEFT n'étant pas subventionnées, elles n'ont pas la possibilité financière d'acquérir des corpus auprès d'agences de redistribution. Cependant de multiples ressources textuelles sont accessibles sur Internet qui constitue un puissant réservoir de données pour l'étude des phénomènes langagiers (Habert et al., 1997). L'aspiration de pages web est réalisable de manière automatique et efficace dès lors que cette procédure respecte un certain nombre de « règles de bonne conduite », permettant l'aspiration des pages sans saturer le serveur.

Pour ce qui concerne les données non récupérables depuis l'Internet, de nombreux échanges sont en général nécessaires. La longueur de la procédure de récupération des données peut être

[9] Respectivement `www.avoir-alire.com` pour les critiques de livres et de films, et `www.jeuxvideos.com` pour les critiques de jeux vidéos.

[10] `http://ntcir.nii.ac.jp/index.php/Table/MOAT/`

due à des réserves liées au caractère juridique de l'accès aux données, ou encore à la surcharge de travail que cela implique souvent pour le dépositaire des données.

2.2 Normalisation des corpus

Formats et encodages L'hétérogénéité des formats sous lesquels se présentent les données récupérées nous impose une phase de normalisation de ces données.

Une phase de nettoyage des corpus est nécessaire, chaque format étant l'objet d'erreurs récurrentes. Par exemple, les pages issues du web peuvent présenter des problèmes de mauvaise structuration, dans le sens où les navigateurs Internet peuvent afficher la page quand bien même une entité HTML aura été mal écrite (par exemple dans le cas de l'oubli du point-virgule final : *&eacute* au lieu de *&eacute ;* pour coder l'accent aigu sur la lettre « e ») ou qu'une balise n'aura pas été fermée (telle une ligne de tableau comprenant la balise ouvrante <*td*> mais pas la balise fermante associée <*/td*>). La conversion en texte brut sera alors imparfaite.

Par ailleurs, les logiciels bureautiques possèdent un encodage propriétaire de certains caractères tels que les points de suspension, les ligatures, ou encore les apostrophes. Il arrive aussi que des fichiers textuels mélangent plusieurs encodages, généralement un encodage principal et un encodage secondaire pour certains caractères non gérés par l'encodage principal ; il est nécessaire dans ce cas de réencoder le document sous un seul encodage.

Dans le cadre de nos campagnes d'évaluation, nous avons jusqu'à présent opté pour l'encodage des documents en ISO Latin 1 (ISO-8859-1) avec explosion de la ligature « œ » en deux caractères « o » et « e ».

Conversion au format XML Nous avons fait le choix de structurer les corpus de la campagne d'évaluation avec XML. Ce format permet d'introduire facilement des méta-données dans les corpus. Les méta-données sont de deux types, d'abord celles décrivant l'origine du corpus et éventuellement les traitements préparatoires, et d'autre part celles liées à la tâche de la campagne, c'est-à-dire pour chaque document, les résultats attendus puis les résultats du test pour chaque participant. Une DTD (Document Type Definition), sorte de grammaire de la structure du document, accompagne chaque corpus.

Ce choix est motivé par deux raisons : en premier lieu, XML est généralement bien reconnu par les outils du TAL, ces derniers étant facilement configurables pour traiter les balises contenant les informations à traiter ; le cas échéant, plusieurs méthodes existent pour traiter efficacement ce type de données (feuille de traitements XSLT ou parseur XML). La seconde raison de ce choix concerne la pérennité des corpus ainsi créés. Lorsqu'aucun problème juridique de redistribution des corpus ne se présente, nous rendons accessibles les corpus créés pour le défi depuis le site `http://deft.limsi.fr/` (portail de base vers les différentes éditions du défi). Le format XML permet un étiquetage sémantique clair et standardisé des différents types de données des corpus, les rendant aisément réutilisables.

En règle générale, nous n'avons réalisé qu'un balisage de base, en l'occurrence l'encadrement de chaque document du corpus par une balise <*doc*>, le texte du document entre balises <*texte*> avec une segmentation au niveau des paragraphes (balises <*p*>) ainsi que les méta-données constituant l'objet de la campagne d'évaluation (valeurs nécessaires à l'apprentissage), en l'occurrence la valeur de l'opinion exprimée dans un texte sur une échelle à 2 ou 3 valeurs dans le cadre de DEFT'07.

2.3 Les corpus

Dans ces corpus, les valeurs d'opinion d'origine ont parfois été regroupées en un nombre restreint de classes. Les raisons de ces choix sont expliquées en section 3.

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, dérivées des cinq classes d'origine : favorable (classe 2 – classes d'origine 3 et 4 regroupées), neutre (classe 1 – classe d'origine 2) et défavorable (classe 0 – classes d'origine 1 et 0 regroupées). Exemples :

- Classe 0 : ... *Mais cette esbroufe formelle ne parvient pas à masquer l'indigence gravissime d'un scénario ... dont l'humour vaseux et les fausses pistes éculées agacent et ennuient ...*
- Classe 1 : ... *Malgré tout, ce roman manque singulièrement d'éclat. Certes, il est jalonné de détails attachants de la vie quotidienne et son écriture est fluide, comme une évidence. Mais la simplicité a ses limites ...*
- Classe 2 : ... *Une jolie comédie qui s'interroge sur les amours de jeunesse, les rêves d'antan et les responsabilités d'aujourd'hui... Une comédie au thème très classique mais menée avec beaucoup de finesse ...*

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, dérivées d'une note d'origine sur vingt : appréciation positive du jeu vidéo (classe 2 – notes d'origine de 15 à 20), appréciation moyenne (classe 1 – notes de 10 à 14) et appréciation négative (classe 0 – notes de 0 à 9). Exemples :

- Classe 0 : *cerveau visiblement réduit qui possède un masque de justicier à la Zorro, ce qui lui donne un look de catcheur d'assez mauvais goût. Ce n'est pas pour rien que cet accessoire s'appelle le masque de la honte dans le jeu ! ... Résultat, on s'ennuie vite aux commandes de ce personnage qui se retrouve au cœur d'une histoire complètement anecdotique, qui n'est là que pour servir de prétexte à une succession de combats toujours identiques ...*
- Classe 1 : ... *En dépit de ces deux gros défauts, Amenophis se laisse suivre de bout en bout sans trop de mal. L'histoire s'enchaîne plutôt bien avec plusieurs cinématiques qui se déclenchent aux moments clés. Les rebondissements sont nombreux (mais prévisibles) et les énigmes, essentiellement basées sur l'observation et la récupération d'objets, ne sont pas insurmontables, bien au contraire ...*
- Classe 2 : ... *Un titre à découvrir absolument par les fans de jeux de rôle. Morrowind est un jeu tout simplement passionnant, d'une richesse incroyable et doté d'une longévité hors-norme. Son principe devrait séduire de nombreux joueurs qui découvriront à cette occasion une expérience de jeu unique ...*

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 confé-

rences JADT[11], RFIA[12] et TALN[13]. 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). Exemples :

- Classe 0 : ... *Hors thème. Il s'agit d'une étude de phonétique acoustique. Cet article pourrait être soumis à une conférence de phonétique ...*
- Classe 1 : ... *Cet article est clair et présente des validations expérimentales sur des corpus réels. Cependant, il n'apporte pas grand-chose de neuf par rapport à la littérature existente - l'originalité et l'aspect novateur sont donc tres faibles.*
- Classe 2 : ... *Il s'agit d'un excellent papier tres bien ecrit et qui apporte beaucoup d'informations interessantes. Les auteurs motivent et proposent un travail d'ingenierie linguistique de haut vol : ils tirent parti de techniques connues en linguistique computationelle pour proposer un systeme finalise effectivement implemente et evalue.*

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[14]. 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). Exemples :

- Classe 0 : ... *L'accès à l'énergie dans des conditions normales et au juste prix ne doit pas devenir le privilège de quelques-uns. Car la privatisation peut être du vol lorsque des actionnaires privés accèdent à vil prix à un patrimoine national comme le réseau de transport du gaz. ...*
- Classe 1 : ... *Nous avons passé des dizaines d'heures en juillet et en août, sous l'autorité du président <hommePolitique />, à analyser ce projet. Mon sentiment est qu'il répond bel et bien à l'évolution du monde : en ce début de siècle, il n'est pas anormal de chercher des solutions nouvelles pour des temps nouveaux. ...*

3 Test humain de la tâche

3.1 Présentation

Lors de la phase de préparation de la campagne, nous avons effectué des tests humains de la tâche. Les juges humains, en l'occurrence les membres du Comité d'organisation du défi, ont eu pour consigne d'effectuer la tâche envisagée pour cette édition du défi sur un échantillon de chaque corpus. Cette participation sur de brefs extraits des corpus vise deux objectifs principaux.

Faisabilité de la tâche Dans un premier temps, la mise en situation des évaluateurs humains sur les tâches envisagées permet de mesurer la faisabilité globale de chacune des tâches. Cette possibilité est évaluée, à la fois en termes des buts poursuivis (la tâche présente t-elle un intérêt

[11] Journées internationales d'Analyse statistique des Données Textuelles.
[12] Reconnaissance des Formes et Intelligence Artificielle.
[13] Traitement Automatique des Langues Naturelles.
[14] L'intégralité des séances de débats sur ce projet de loi est accessible à l'adresse `http://www.assemblee-nationale.fr/12/debats/`

suffisant ?), en termes de difficulté de la tâche (est-elle facile ou difficile pour des humains ?), et enfin en termes de disponibilité et de richesse des données accessibles (les corpus rassemblent-ils suffisament d'informations pour mener à bien cette tâche ?). Par la participation des juges humains, nous mesurons l'intérêt des pistes envisagées et nous nous faisons une première idée des résultats auxquels il est possible de prétendre pour ces pistes.

Création et adaptation des données de référence Dans le cadre des différentes éditions du défi DEFT, nous construisons automatiquement les données de référence à partir des méta-données dont nous disposons pour chacun des documents de chaque corpus.

Pour ce qui concerne l'édition 2007 du défi, nous avons rassemblé des corpus pour lesquels existait une correspondance entre chaque texte d'opinion et une note résumant l'opinion détaillée dans le texte (avis général pour les relectures d'articles scientifiques, note globale pour chaque critique de livre, de film, de jeux vidéos, et valeur du vote de chaque parlementaire s'exprimant). La note globale assignée à chaque texte constituait, pour cette édition, la référence. Chacun de ces corpus comprenait ainsi une échelle de valeurs qui lui était spécifique (2 valeurs pour les débats parlementaires, 4 valeurs pour les relectures d'articles, 5 valeurs pour les critiques de livres et de films, et 20 valeurs pour les critiques de jeux vidéos).

Le second objectif du test humain de la tâche a donc été de définir précisément les échelles de valeurs à utiliser dans le cadre de la tâche. Il était possible de mettre en œuvre deux solutions : soit conserver les échelles d'origine avec les variantes que cela impose pour chaque corpus, soit essayer de réduire et d'homogénéiser les différentes échelles. Afin de prendre cette décision, les évaluateurs humains ont dû, pour chaque texte de chaque corpus, attribuer une valeur sur l'échelle d'origine et sur l'échelle restreinte.

Les résultats ont été évalués en terme de F-mesure (voir la section 4.1 pour la définition de la F-mesure), et les accords entre juges mesurés par le coefficient Kappa (Cohen, 1960; Carletta, 1996). Ce coefficient permet de mettre en évidence un taux d'accord entre deux juges (P_0 correspond à la proportion d'accords observée et P_e à la proportion d'accords aléatoire, ou concordance attendue sous l'hypothèse d'indépendance des jugements). Il est d'autant plus faible que la proportion d'accords observée se rapproche de celle qui serait obtenue de façon aléatoire. Il est égal à 1 lorsque l'accord entre juges est parfait.

$$\text{Kappa} = \frac{P_0 - P_e}{1 - P_e}$$

L'accord entre juges est qualifié selon la valeur prise par le coefficient κ. Il est qualifié d'excellent pour un κ compris entre 0,81 et 1,00, il est bon entre 0,61 et 0,80, modéré entre 0,41 et 0,60, médiocre entre 0,21 et 0,40, mauvais entre 0 et 0,20 et l'accord est jugé très mauvais si le κ est négatif.

Dans notre cas de classification de textes d'opinion, nous voulions d'abord tester la difficulté de la tâche, mais aussi les accords entre juges sur les valeurs d'opinion présentes dans les corpus. Sur la base des accords ainsi révélés, il a été rendu possible de sélectionner l'échelle de valeurs à utiliser pour la constitution de la référence de chaque corpus.

3.2 Les valeurs d'opinion, échelle large ou échelle restreinte

Les quatre corpus retenus pour la campagne présentaient chacun la particularité de combiner une note ou une opinion avec un texte descriptif, la note venant résumer le jugement

exprimé dans le texte. En raison des sources différentes utilisées, nous avons été confrontés à autant d'échelles de valeurs qu'il y avait de corpus :

- 2 valeurs pour les débats parlementaires (le parlementaire s'exprimant dans ces débats était, soit favorable à la loi en examen, soit défavorable) ;
- 4 valeurs pour les relectures d'articles scientifiques (*accepté en l'état – accepté avec modifications mineures – accepté avec modifications majeures et seconde évaluation complète – rejeté*) ;
- 5 valeurs pour les critiques de livres et de films (une note comprise entre 0 et 4) ;
- 20 valeurs pour les critiques de jeux vidéos (en l'occurrence une note telle que celles utilisées dans le système éducatif français).

Les juges humains avaient pour consigne d'attribuer à chaque document une valeur d'opinion sur deux types d'échelles, l'échelle large conservant les valeurs d'origine, et une échelle restreinte ramenant le nombre de ces valeurs à 3 maximum.

Le corpus des critiques de jeux vidéo Les Tableaux 1 donnent les coefficients κ obtenus par trois 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 pour les notes de 0 à 9, classe 1 pour celles de 10 à 14 et classe 2 pour celles de 15 à 20. Ce changement d'échelle a été choisi après une recherche d'accord intra-juge, c'est-à-dire en comparant les résultats d'un même juge d'une part suivant une échelle restreinte, et d'autre part suivant l'échelle large en appliquant ensuite la conversion d'échelle à ces résultats.

Juge	Réf.	1	2	3
Réf.		0,17	0,12	0.07
1	0,17		0,03	0.05
2	0,12	0,03		0.07
3	0.07	0.05	0.07	

Juge	Réf.	1	2	3
Réf.		0,74	0,79	0.69
1	0,74		0,74	0.54
2	0,79	0,74		0.69
3	0.69	0.54	0.69	

TAB. 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).*

Ces résultats montrent un mauvais accord entre les juges ainsi qu'entre juge et référence, 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).

Juge	1	2	3
F-mesure	0,13	0,16	0.06

Juge	1	2	3
F-mesure	0,86	0,90	0.73

TAB. 2 – *F-mesures des résultats des juges humains 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 résultats des juges en termes de F-mesure confirment ce résultat. On le voit clairement sur les deux Tableaux 2 qui donnent les F-mesures obtenues par chaque juge sur ce corpus, avec les deux échelles de notes. Lorsqu'il s'agit d'interpréter le degré de jugement favorable ou défavorable exprimé par un texte, une échelle restreinte de valeurs d'opinion donne une interprétation plus fiable.

Le corpus des critiques de films et livres L'échelle des valeurs d'opinion associées originellement à ce corpus est moins large que celle des jeux vidéos. Traditionnellement, l'échelle de jugement comporte 5 valeurs : les films sont jugés de très mauvais à excellent en passant par mauvais, moyen, et bon. Nous avons constaté une amélioration des résultats des juges humains, ainsi que de leur degré d'accord, par le changement d'échelle vers seulement trois valeurs d'opinion : mauvais (pour mauvais et très mauvais), moyen, bon (pour bon et excellent).

Le Tableau 3 montre les degrés d'accord entre cinq juges sur ce corpus. Les accords entre juges humains se sont révélés mauvais à modérés pour l'échelle large tandis que ces accords deviennent médiocres à bons dans le cas d'une échelle restreinte à trois valeurs (Tableau 3). Nous voyons que les différences entre les jugements utilisant l'une et l'autre des deux échelles sont beaucoup moins accentuées que dans le cas du corpus des jeux vidéos. C'est tout à fait cohérent avec le fait que les deux échelles de valeurs d'opinion sont beaucoup plus proches l'une de l'autre.

Juge	1	2	3	4	5
1		0.37	0.49	0.48	0.35
2	0.37		0.36	0.30	0.43
3	0.49	0.36		0.49	0.54
4	0.48	0.30	0.49		0.60
5	0.35	0.43	0.54	0.60	

Juge	1	2	3	4	5
1		0.45	0.43	0.57	0.37
2	0.45		0.73	0.48	0.54
3	0.43	0.73		0.62	0.62
4	0.57	0.48	0.62		0.76
5	0.37	0.54	0.62	0.76	

TAB. 3 – *Coefficient κ entre 5 juges humains sur le corpus des films. Échelle de valeurs d'opinion de 0 à 4 (tableau de gauche) et de 0 à 2 (tableau de droite).*

Le Tableau 4 montre les résultats des juges en termes de F-mesure. Le changement d'échelle améliore aussi leurs scores, mais toujours moins nettement que pour le corpus des jeux vidéo.

Juge	1	2	3	4	5
F	0.29	0.43	0.54	0.59	0.61

Juge	1	2	3	4	5
F	0.52	0.76	0.69	0.70	0.79

TAB. 4 – *F-mesures des résultats de 5 juges humains sur le corpus des films. Échelle de valeurs d'opinion de 0 à 4 (tableau de gauche) et de 0 à 2 (tableau de droite).*

Le mauvais accord entre juges sur les échelles larges montre la difficulté à la fois d'exprimer et d'interpréter précisément des nuances d'opinion. En effet, sur une échelle large deux valeurs voisines deviennent trop proches, leurs différences sont peu marquées.

Les résultats de ces différents tests nous ont donc amenés à utiliser des échelles restreintes dans le cadre de cette campagne, en proposant des échelles d'opinion comprenant 2 ou 3 va-

leurs différentes selon les corpus concernés. Il nous a semblé plus réaliste de comparer les résultats des logiciels à des données de référence qui se sont révélées robustes, en cela qu'elles suivent une échelle de valeurs d'opinion sur laquelle les juges humains sont en meilleur accord.

4 Mesures d'évaluation

La tâche à effectuer par les équipes participantes consistait à attribuer à chaque document de chaque corpus une classe d'opinion, prise parmi 2 ou 3 classes prédéfinies. Nous détaillons dans les sections qui suivent les mesures choisies pour évaluer les résultats des participants. Ces mesures sont couramment utilisées pour évaluer des classifieurs (Nakache et Métais, 2005).

4.1 Les mesures

Rappel et précision Le rappel et la précision mettent en avant la proportion de résultats corrects obtenus, d'une part par rapport au nombre attendu de résultats corrects, et d'autre part par rapport au nombre total de résultats obtenus.

Soient :

– a = nombre de documents sélectionnés par le système et qui sont pertinents
– b = nombre de documents sélectionnés par le système mais qui ne sont pas pertinents
– c = nombre de documents non sélectionnés par le système mais qui sont pertinents

On a :

$$\text{Précision} = \frac{a}{a+b} \qquad \text{Rappel} = \frac{a}{a+c}$$

Classification à n classes Lorsque le rappel et la précision sont utilisés pour évaluer la performance d'un algorithme de classification à n classes, les moyennes globales de la précision et du rappel sur l'ensemble des n classes peuvent être évaluées de 2 manières (Sebastiani, 2005). La micro-moyenne fait d'abord la somme des éléments du calcul (a, b, et c) sur l'ensemble des n classes, pour calculer la précision et le rappel globaux. En revanche, la macro-moyenne 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 d'effectif important comptera donc davantage dans la moyenne qu'une classe de faible effectif. Dans ce cas, l'évaluation de la performance d'un classifieur, sur un corpus comportant des classes d'effectifs très différents, sera très dépendante de la performance de la classe de plus fort effectif.

En revanche, dans la macro-moyenne, le score pour chaque classe d'un corpus compte à égalité avec les scores des autres classes, quels que soient leurs effectifs respectifs. L'évaluation de la performance d'un classifieur sur un corpus dépendra alors autant des classes de faible effectif que des classes d'effectif important.

Soient :

– a_i = nombre de documents correctement attribués à la classe i ;
– b_i = nombre de documents faussement attribués à la classe i ;
– c_i = nombre de documents appartenant à la classe i mais non attribués à la classe i ;

– n = nombre de classes.

Micro-moyennes :

$$\text{Précision} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i} \quad \text{Rappel} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} c_i}$$

Macro-moyennes :

$$\text{Précision} = \frac{\sum_{i=1}^{n} \left(\frac{a_i}{(a_i + b_i)} \right)}{n} \quad \text{Rappel} = \frac{\sum_{i=1}^{n} \left(\frac{a_i}{(a_i + c_i)} \right)}{n}$$

La F-mesure Le rappel et la précision donnent deux points de vue différents sur les résultats d'un test. La F-mesure (ou F-score) a été introduite par (van Rijsbergen, 1979) pour obtenir une mesure unique privilégiant l'un ou l'autre des deux points de vue suivant la valeur affectée au paramètre β.

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

Lorsque $\beta = 1$, la F-mesure est la moyenne harmonique entre la précision et le rappel. C'est la mesure que nous avons utilisée, elle privilégie un équilibre entre le rappel et la précision.

Lorsque la F-mesure est utilisée pour évaluer la performance d'une classification à n classes, on peut utiliser soit la micro-moyenne des précisions et rappels, soit leur macro-moyenne. Les classes d'opinion étant inégalement réparties dans les corpus, nous avons choisi de calculer la F-mesure globale avec la macro-moyenne pour que les résultats sur chaque classe comptent de la même manière quelle que soit la taille de la classe.

4.2 Les indices de confiance

Un système de classification automatique peut attribuer à un document une distribution de probabilités sur les différentes classes au lieu de lui attribuer une seule classe. Pour chaque document, l'indice de confiance attribué à une classe est la probabilité que ce document appartienne à cette classe. Pour chaque document, la somme des indices de confiance sur toutes les classes est donc égale à 1. Par ailleurs, les notions de document *correctement* ou *faussement* attribué à une classe ne sont plus pertinentes dans ce cas, un document étant attribué à l'une ou l'autre des classes avec une probabilité plus ou moins forte (éventuellement nulle). Il est clair néanmoins que si un document est attribué avec la plus forte probabilité à une classe non pertinente pour ce document, il sera, si on ne sélectionne que la classe de plus forte probabilité, *faussement* attribué à cette classe.

La F-mesure pondérée par l'indice de confiance a été utilisée à titre indicatif pour des comparaisons complémentaires entre les méthodes mises en place par les équipes.

Dans la F-mesure pondérée, la précision et le rappel pour chaque classe i sont pondérés par l'indice de confiance. Ce qui donne :

$$\text{Précision}_i = \frac{\sum_{d \in D_i} IC_d(i)}{\sum_{d=1}^{N} IC_d(i)}$$

$$\text{Rappel}_i = \frac{\sum_{d \in D_i} IC_d(i)}{N_i}$$

Avec :
- D_i : ensemble des documents appartenant à la classe i (documents pertinents pour la classe i) ;
- N_i : nombre de documents appartenant à la classe i (=card(D_i)) ;
- N : nombre total de documents (somme des N_i sur l'ensemble des classes, celles-ci formant une partition sur les documents) ;
- $IC_d(i)$: indice de confiance attribué par le système à la classe i pour le document d.

La F-mesure pondérée est ensuite calculée à l'aide des formules de la F-mesure classique (voir section 4.1).

5 Résultats de la campagne d'évaluation

5.1 F-mesure

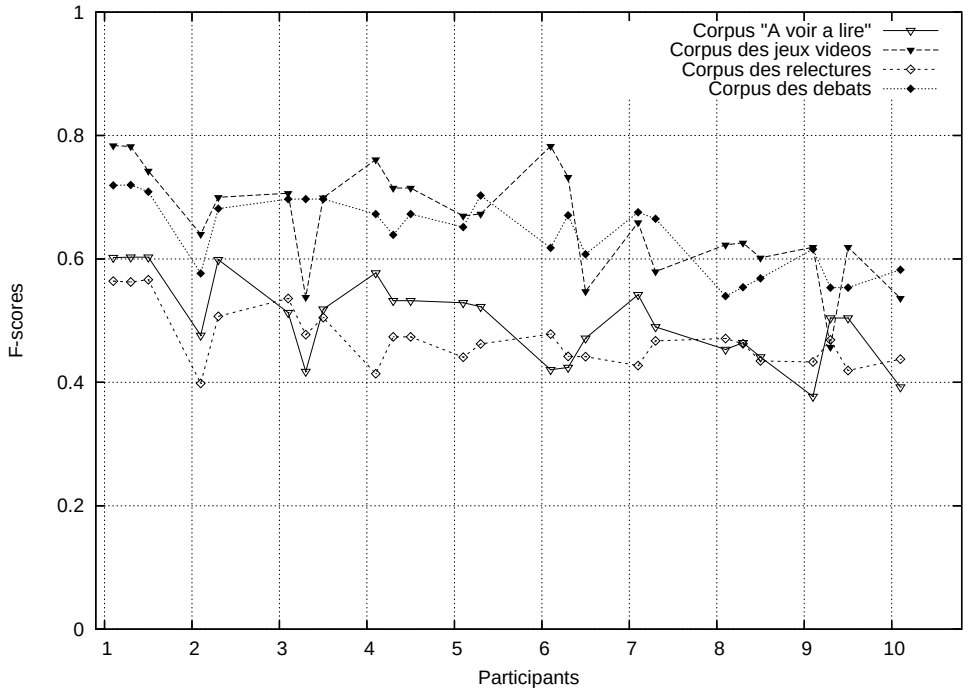

FIG. 1 – *F-mesure classique ($\beta = 1$) pour l'ensemble des soumissions de chacun des candidats.*

Équipe	Soumission	À voir, à lire	Jeux vidéos	Relectures	Débats
Torres-Moreno et al. (2007)	1	0.602	**0.784**	0.564	0.719
Torres-Moreno et al. (2007)	2	**0.603**	0.782	0.563	**0.720**
Torres-Moreno et al. (2007)	3	**0.603**	0.743	**0.566**	0.709
Ahat et al. (2007)	4	0.476	0.640	0.398	0.577
Ahat et al. (2007)	5	0.599	0.699	0.507	0.681
Maurel et al. (2007)	6	0.513	0.706	0.536	0.697
Maurel et al. (2007)	7	0.418	0.538	0.477	0.697
Maurel et al. (2007)	8	0.519	0.700	0.505	0.697
Vernier et al. (2007)	9	0.577	0.761	0.414	0.673
Vernier et al. (2007)	10	0.532	0.715	0.474	0.639
Vernier et al. (2007)	11	0.532	0.715	0.474	0.673
Crestan et Acuna-Agost (2007)	12	0.529	0.670	0.441	0.652
Crestan et Acuna-Agost (2007)	13	0.523	0.673	0.462	0.703
Plantié et al. (2007)	14	0.421	0.783	0.478	0.618
Plantié et al. (2007)	15	0.424	0.732	0.442	0.671
Plantié et al. (2007)	16	0.472	0.547	0.442	0.608
Trinh (2007)	17	0.542	0.659	0.427	0.676
Trinh (2007)	18	0.490	0.580	0.467	0.665
Généreux et Santini (2007)	19	0.453	0.623	0.471	0.540
Généreux et Santini (2007)	20	0.464	0.626	0.463	0.554
Généreux et Santini (2007)	21	0.441	0.602	0.435	0.569
Charton et Acuna-Agost (2007)	22	0.377	0.619	0.433	0.616
Charton et Acuna-Agost (2007)	23	0.504	0.457	0.469	0.553
Charton et Acuna-Agost (2007)	24	0.504	0.619	0.419	0.553
Acosta et Bittar (2007)	25	0.392	0.536	0.437	0.582

TAB. 5 – *F-mesures pour toutes les soumissions sur chaque corpus.*

Pour cette édition du défi, chaque candidat avait la possibilité de soumettre jusqu'à trois résultats pour chacun des corpus. Nous avons eu 25 soumissions, et nous avons donc évalué 25 fichiers de résultats pour chaque corpus. Pour chaque corpus de chaque soumission, nous avons calculé la F-mesure classique[15], celle calculée sans les indices de confiance, et la F-mesure pondérée par les indices de confiance.

Le Tableau 5 présente les résultats par la F-mesure classique dans l'ordre du classement des équipes suivant leur meilleure soumission. Pour chaque corpus le meilleur score est marqué en gras.

Au regard de ces résultats, il apparaît assez nettement que les quatre corpus ont posé des problèmes distincts dans les traitements mis en œuvre. Nous pouvons ainsi établir un classement des corpus sur la base des F-mesures obtenues, ces résultats traduisant leur plus ou moins grande difficulté de traitement.

[15]Nous l'avons aussi appelée F-mesure stricte pour la différencier de la F-mesure pondérée.

Classement des corpus par difficulté de traitement On observe les résultats les meilleurs pour les corpus des jeux vidéos et des débats parlementaires et, à l'inverse, de moins bons résultats pour les corpus des critiques de films et les relectures. Cette tendance semble partagée par l'ensemble des participants au défi comme l'atteste le graphique de la Figure 1[16].

1. Corpus des jeux vidéos : F-mesures comprises entre 0,784 et 0,457 ;
2. Corpus des débats parlementaires : F-mesures comprises entre 0,720 et 0,540 ;
3. Corpus « À voir, à lire » : F-mesures comprises entre 0,602 et 0,377 ;
4. Corpus des relectures : F-mesures comprises entre 0,566 et 0,398.

Outre le fait que cette gradation de la difficulté des différents corpus apparaît partagée par l'ensemble des participants, les Tableaux 2 et 4 (section 3.2) montrent que ces résultats rejoignent les tests effectués par les juges humains :

1. Corpus des débats parlementaires : F-mesures comprises entre 0,90 et 0,79 ;
2. Corpus des jeux vidéos : F-mesures comprises entre 0,90 et 0,73 ;
3. Corpus « À voir, à lire » : F-mesures comprises entre 0,79 et 0,52 ;
4. Corpus des relectures : F-mesures comprises entre 0,58 et 0,41.

Les évaluateurs humains ont obtenu de meilleurs résultats sur les corpus des jeux vidéos et « *à voir, à lire* » que les systèmes automatiques des participants au défi. En revanche, les résultats sont quasi-identiques entre juges humains et systèmes automatiques sur le corpus des relectures, corpus jugé difficile à noter par les humains.

Les caractéristiques des corpus Si on met en parallèle avec ce classement les caractéristiques des différents corpus, on peut remarquer que le corpus des jeux vidéos comporte des critiques longues, argumentées et très structurées, avec systématiquement une description générale suivie de commentaires plus spécifiques sur un ensemble de sous-thèmes tels que le graphisme, la jouabilité, la durée de vie, la bande son et le scénario, et enfin une conclusion évaluative. Chaque document fait en moyenne 7 ko, contre 2,5 ko pour les critiques de films et livres et 2,4 ko pour les relectures d'articles. La taille moyenne de chaque intervention dans les débats parlementaires est encore plus courte (1,3 ko), en revanche ce corpus ne comporte que deux classes d'opinion.

Disparités entre les résultats La Figure 1 met en relief les comportements des différents logiciels face à ces différences entre les corpus. Nous constatons que la ligne de partage de la difficulté reste stable entre d'une part les corpus des débats parlementaires et des jeux vidéo, et d'autre part les corpus des relectures et des films et livres, à une exception près, celle de la soumission 23 (équipe 9) qui a le plus mauvais score sur le corpus des jeux vidéos. En revanche, les lignes de résultats se croisent parfois entre le corpus des jeux vidéos et celui des débats, et entre le corpus des relectures et celui des films et livres. En effet, malgré les difficultés rencontrées généralement sur le corpus des relectures, dans quelques cas, les participants semblent avoir eu moins de difficultés pour ce corpus que pour celui des critiques de livres et de films.

[16] Les points figurant les différentes exécutions des participants sur le même corpus ont été reliés pour une meilleure visibilité des performances globales par corpus.

Il en est ainsi pour les soumissions 6, 7, 14, 15, 19, 22 et 25 (voir le Tableau 5). Par ailleurs, si l'on considère les équipes jeunes chercheurs indépendamment des autres équipes (équipes 7, 9 et 10 dans le Tableau 5), une singularité émerge quant au corpus des débats parlementaires. Alors que les meilleurs résultats ont été obtenus sur le corpus des jeux vidéos, les équipes de jeunes chercheurs ont obtenu leurs meilleurs résultats sur le corpus des débats parlementaires. Le plus grand écart de résultat entre les corpus est atteint par la soumission 14 (équipe 6), avec d'une part un score de 0.783 pour les jeux vidéos, soit presque le meilleur score (0.784), et d'autre part un score de 0.421 pour les critiques de films et livres, qui figure parmi les mauvais scores sur ce corpus.

5.2 Indices de confiance

Les participants ont eu la possibilité d'associer un indice de confiance à chaque note attribuée aux documents des corpus. Cet indice de confiance était proposé de manière optionnelle.

Sur les dix participants au défi, six y ont recouru. Sur ces six participants, certains l'ont appliqué pour chaque soumission, d'autres n'ont proposé que certaines soumissions avec indice de confiance. Le tableau 6 donne les résultats des soumissions avec les scores de confiance. On a rappelé entre parenthèses les F-mesures classiques. Les meilleurs scores sont indiqués en gras.

Équipe	Soumission	À voir, à lire	Jeux vidéos	Relectures	Débats
1	1	**0.525** (0.602)	0.682 (**0.784)**	0.490 (0.564)	0.648 (0.719)
1	2	**0.525** (**0.603)**	0.682 (0.782)	0.490 (0.563)	0.649 (**0.720)**
1	3	0.511 (**0.603)**	0.633 (0.743)	0.467 (**0.566)**	0.629 (0.709)
3	6	0.514 (0.513)	**0.701** (0.706)	0.522 (0.536)	**0.695** (0.697)
3	7	0.417 (0.418)	0.539 (0.538)	0.479 (0.477)	**0.695** (0.697)
3	8	0.519 (0.519)	0.699 (0.700)	0.504 (0.505)	**0.695** (0.697)
4	9	0.386 (0.577)	0.459 (0.761)	0.377 (0.414)	0.592 (0.673)
6	14	0.387 (0.421)	0.624 (0.783)	0.382 (0.478)	0.547 (0.618)
6	16	0.447 (0.472)	0.524 (0.547)	0.367 (0.442)	0.594 (0.608)
7	17	0.456 (0.542)	0.567 (0.659)	0.356 (0.427)	0.596 (0.676)
7	18	0.368 (0.490)	0.381 (0.580)	0.368 (0.467)	0.546 (0.665)
8	19	0.392 (0.453)	0.519 (0.623)	0.450 (0.471)	0.519 (0.540)
8	20	0.394 (0.464)	0.518 (0.626)	**0.524** (0.463)	0.403 (0.554)
8	21	0.371 (0.441)	0.519 (0.602)	0.431 (0.435)	0.520 (0.569)

TAB. 6 – *F-mesures pondérées, avec entre parenthèses la F-mesure classique, pour les soumissions ayant utilisé les indices de confiance. La correspondance entre numéro d'équipe et articles est la suivante : 1 – Torres-Moreno et al. (2007), 3 – Maurel et al. (2007), 4 – Vernier et al. (2007), 6 – Plantié et al. (2007), 7 – Trinh (2007), 8 – Généreux et Santini (2007).*

Les indices de confiance ont un effet de lissage sur les résultats. Globalement, on obtient une baisse des meilleurs résultats et une hausse des plus faibles. Des indices de confiance presque semblables pour un même document en feront monter le score, même si le plus élevé d'entre eux n'est pas attribué à la valeur correcte d'opinion. Les meilleurs résultats suivant

la F-mesure pondérée ne correspondent donc pas aux meilleurs résultats suivant la F-mesure classique. Cependant, certains participants déclarent qu'ils ont, dans un premier temps, mis en compétition plusieurs classifieurs dont ils disposaient, et conservé ceux qui fournissaient des résultats bien tranchés, de préférence à ceux qui se montraient perplexes. C'est donc lors de cette étape de comparaison préalable que l'indice de confiance a montré pour eux son utilité.

6 Bilan des méthodes de classification de textes d'opinion

Les méthodes statistiques avec apprentissage ont été largement utilisées avec des résultats parfois bien différents. La différence tient essentiellement dans une bonne utilisation d'une grande variété de méthodes pour représenter les textes et ensuite pour les classer. L'équipe arrivée première (Torres-Moreno et al., 2007) a ainsi utilisé un grand nombre de paramètres de représentation du texte – vocabulaire d'opinion, expressions repérées par des règles, règles de ré-écriture de noms propres, etc – et six classifieurs différents avec fusion des résultats.

En revanche, l'équipe arrivée deuxième (Ahat et al., 2007) n'a utilisé qu'une méthode, l'analyse sémantique latente, et a été aussi la seule à l'utiliser. Les principaux tests ont porté sur la séparation des corpus pour l'apprentissage.

Une approche encore différente, mixte, statistique et symbolique, a été développée par l'équipe arrivée troisième (Maurel et al., 2007). La méthode symbolique est basée sur une analyse syntaxique des textes et une grammaire de relations d'opinion. Le classifieur statistique (Bayes naïf ou SVM avec des résultats presque semblables) est entraîné sur les phrases jugées subjectives pour chaque document par la méthode symbolique.

Une grande variété de méthodes ont donc été utilisées par l'ensemble des participants. Dans ce qui suit, nous les passons en revue car il nous semble qu'elles contiennent toutes des éléments intéressants.

6.1 Représentation du texte

Les participants au défi ont généralement traité les corpus en deux étapes : une première étape de représentation du texte suivie d'une seconde étape de classification. La première étape conduit à sélectionner les traits qui vont représenter le texte. Cette sélection peut être plus ou moins élaborée, mais son but est toujours une réduction, parfois drastique, de l'ensemble des traits pouvant représenter les textes. Elle vise à ne retenir du texte que les caractéristiques les plus pertinentes au regard de l'ensemble prédéfini des classes d'opinion. Différentes approches ont été mises en œuvre par les équipes participantes, utilisées de façon séparée ou complémentaire. Une première approche repose sur l'identification de passages du texte jugés pertinents, parce que subjectifs, c'est-à-dire porteurs d'une opinion ou d'un sentiment. Une autre approche consiste à extraire du texte des termes d'opinion d'après un vocabulaire répertorié. Enfin des approches statistiques plus traditionnelles ont aussi été utilisées.

Passages pertinents Certaines équipes ont choisi de ne retenir qu'une partie du texte : les segments qui leur paraissaient pertinents pour l'évaluation de l'opinion. Ces segments peuvent être prédéfinis, comme l'introduction ou la conclusion d'un texte, ou bien extraits par des méthodes linguistiques.

Les parties de texte prédéfinies ont été soit le premier et le dernier paragraphe du texte, ou, lorsque le texte ne comportait qu'un paragraphe, les deux premières et les deux dernières phrases. Plusieurs équipes ont tenu compte plus spécifiquement de ces parties de texte (Acosta et Bittar, 2007; Charton et Acuna-Agost, 2007; Vernier et al., 2007).

L'extraction de segments très courts de textes a été réalisée suivant deux méthodes linguistiques différentes. Maurel et al. (2007) ont mis en œuvre une méthode d'extraction de relations d'opinion à partir d'une analyse syntaxique de la phrase. La relation de sentiment s'établit autour de l'expression du sentiment et de la cause du sentiment. La grammaire construite pour les détecter prend en compte un vocabulaire d'opinion et les termes du domaine traité. Les auteurs ont ainsi défini des listes de termes propres à chaque thématique traitée dans les corpus : une thématique livres/films pour le corpus « à voir à lire » (*film, livre, album,* etc.), une thématique jeux vidéos pour le corpus « jeux vidéos » (*jeu, graphisme, soft,* etc.) et une thématique articles pour le corpus « relectures » (*article, papier, résultat,* etc). Dans le même esprit, Charton et Acuna-Agost (2007) et Vernier et al. (2007) ont testé des méthodes d'extraction du segment de texte autour d'un mot attracteur, l'un des termes du domaine traité par le corpus.

Vocabulaire d'opinion Plusieurs équipes ont utilisé un vocabulaire d'opinion, suivant des méthodes différentes. Il existe plusieurs ressources linguistiques permettant de classer le vocabulaire selon l'opinion qu'il exprime. Généreux et Santini (2007) ont ainsi mobilisé plusieurs de ces ressources en complémentarité.

WordNet-Affect est un sous-groupe de WordNet qui distingue le vocabulaire selon trois classes : le vocabulaire positif (*joie, rayonné, exalté, allègrement,* etc.), négatif (*crainte, effrayé, terrible, alarme,* etc.) et neutre (*apathie, impassibilité, rêveur, langoureusement,* etc).

Big-Six (Ekman (1972) cité par Généreux et Santini (2007)) classe le vocabulaire en six grandes classes qui correspondent aux six émotions de base ressenties dans le cadre d'expériences en psychologie réalisées dans les années 70 : « colère » (*ombrage, offense, folie, irritation,* etc.), « joie » (*culte, adoration, chaleur, triomphe,* etc.), « tristesse » (*ennui, poids, apitoiement, douleur,* etc.), « dégoût » (*répugnance, horreur, nausée, maladie,* etc.), « peur » (*agitation, effroi, cercueil, timidité,* etc.), et « surprise » (*admiration, étonnement, stupeur, terreur,* etc).

SentiWordNet (Esuli et Sebastiani, 2006) repose sur le lexique de WordNet et attribue trois scores à chaque terme du vocabulaire : un score positif, un score négatif et un score neutre (*adoration :* positif=0.625, négatif=0.000, neutre=0.375 – *affligé :* positif=0.125, négatif=0.625, neutre=0.250 – *affreux :* positif=0.000, négatif=0.625, neutre=0.375, etc).

Les quatorze facettes linguistiques fonctionnelles ont été définies par Santini (2007) (citée par Généreux et Santini (2007)) pour identifier automatiquement le genre des pages web. Ces facettes distinguent les pronoms de personnes, les prédicats, les éléments nominaux et sept catégories de verbes : « verbes d'activité » (*faire, saisir, aller, donner,* etc.), « verbes de communication » (*dire, raconter, appeller, demander,* etc.), « verbes mentaux » (*voir, savoir, penser, trouver,* etc.), « verbes causatifs » (*aider, laisser, forcer, exiger,* etc.), « verbes d'occurrence » (*devenir, survenir, changer, mourir,* etc.), « verbes existentiels » (*devoir, apparaître, tenir, rester,* etc.), et les « verbes aspectuels » (*devoir, apparaître, tenir, rester,* etc).

Certains participants ont créé manuellement une liste de termes positifs et négatifs, en complément de ces ressources préexistantes (Torres-Moreno et al., 2007).

Méthodes statistiques de sélection des traits La majorité des participants au défi a utilisé des méthodes statistiques, seules ou en complément des méthodes linguistiques, pour discriminer les traits importants de chaque texte. Ces méthodes statistiques reposent sur des mesures classiques telles que le tf*idf, le gain d'information, l'information mutuelle, ou encore le critère d'impureté de Gini (Crestan et Acuna-Agost, 2007).

La construction de n-grammes de mots (Vernier et al., 2007), et le calcul de collocations (Torres-Moreno et al., 2007) ont été utilisés pour la détection d'expressions typiques.

Ahat et al. (2007) ont construit des concepts par analyse sémantique latente. Cette technique consiste à extraire les relations entre mots en fonction des occurrences communes de ces mots dans les textes. La matrice de fréquences ainsi obtenue est alors décomposée en valeurs singulières (DVS) de manière à réduire la taille de la matrice. Le résultat obtenu correspond à un espace sémantique à partir duquel il va être possible de comparer les vecteurs de mots au moyen d'indices de similarité (distance euclidienne ou cosinus de l'angle). Une seule équipe a utilisé cette technique, qui s'est révélée très efficace.

6.2 Classification

Les classifieurs Le classifieur le plus utilisé (Trinh, 2007) a été la machine à vecteur de support, SVM (Vapnik, 1995), mais ce n'est pas celui qui a produit les meilleurs résultats. En effet, ce classifieur est très sensible au sur-apprentissage, ce qui le rend moins robuste que d'autres sur les corpus de tests.

Une autre méthode utilisée plusieurs fois avec un certain succès a été la sommation de scores calculés sur chaque terme d'un document (Crestan et Acuna-Agost, 2007; Charton et Acuna-Agost, 2007; Vernier et al., 2007), ou sur chaque relation d'opinion (Maurel et al., 2007).

Parmi les autres méthodes de classification on trouve les arbres de décision (Acosta et Bittar, 2007; Torres-Moreno et al., 2007; Plantié et al., 2007), la régression logistique (Charton et Acuna-Agost, 2007; Acosta et Bittar, 2007), des méthodes probabilistes (Torres-Moreno et al., 2007; Plantié et al., 2007; Maurel et al., 2007), des réseaux de neurones (Plantié et al., 2007), un algorithme de boosting (Torres-Moreno et al., 2007), l'algorithme des k plus proches voisins (Torres-Moreno et al., 2007), un classifieur à base de règles d'association (Vernier et al., 2007), et un calcul de similarité entre le vecteur représentant une classe et le vecteur représentant un texte (Charton et Acuna-Agost, 2007; Ahat et al., 2007).

Méthodes hybrides Certaines équipes ont conçu des méthodes hybrides utilisant au moins deux classifieurs (Torres-Moreno et al., 2007; Plantié et al., 2007; Maurel et al., 2007). Ce sont Torres-Moreno et al. (2007) qui ont poussé le plus loin la méthode en prenant 6 classifieurs avec des variantes dans la représentation du texte donnant 9 systèmes de décision, un même poids étant attribué à chaque système dans la fusion finale. Cette méthode a produit les meilleurs scores.

7 Conclusion et perspectives

Depuis quelques années, les travaux de classifications de documents prennent en compte l'opinion exprimée dans les documents. Pang et Lee (2008) considèrent que cette détection

repose aussi bien sur l'identification des informations subjectives contenues dans les textes que sur la mise en évidence des portions de documents qui contiennent ces informations.

Ce nouveau niveau d'analyse des textes fait également l'objet de pistes dédiées dans le cadre de campagnes d'évaluation. Les campagnes TREC ont mis en place en 2006 une piste sur l'analyse des blogs. L'objectif visé concernait la recherche d'information sur la blogosphère. L'une des tâches concernait plus spécifiquement la détection des opinions exprimées sur un thème donné (Ounis et al., 2006).

L'étendue et la variété des approches utilisées par les participants à DEFT'07 nous ont semblé remarquables. La sélection des traits représentant le texte est une étape importante dans une tâche de classification. De ce point de vue, l'utilisation d'un vocabulaire d'opinion, plus particulièrement en association avec l'utilisation des termes du domaine spécifiques à chaque corpus, a donné de bons résultats. Par ailleurs, les méthodes hybrides de classification, mettant en concurrence plusieurs classifieurs, se sont révélées performantes. Les résultats semblent améliorés par l'utilisation de méthodes complémentaires ou par un vote sur plusieurs méthodes.

Il apparaît que les modèles ayant produit de meilleurs résultats sur ces corpus d'opinion sont ceux qui reposent sur une combinaison de méthodes – en particulier les méthodes probabilistes – et dont la représentation du texte se fonde sur l'ensemble du texte (et non une sélection de passages jugés pertinents) avec la prise en compte du vocabulaire d'opinion. À l'inverse, les modèles fonctionnant sur des extraits de textes, sans recourir à un vocabulaire d'opinion ou à une sélection des traits importants de chaque texte, et ne s'appuyant que sur une seule méthode, semblent ne pas être adaptés au regard des résultats obtenus. Notons qu'il est cependant difficile de définir avec certitude quel modèle correspond le mieux à une tâche précise, étant donnée que certains paramètres utilisés par les participants n'ont pas été détaillés dans les articles.

Les résultats de cette campagne ont aussi donné des indications intéressantes sur les corpus fournis aux participants. En effet, les résultats des tests faits sur de courts extraits des différents corpus par des juges humains, montrent le même ordre de difficulté que les méthodes automatiques : le corpus des relectures a été le plus difficile à évaluer, et celui des tests des jeux vidéos le plus facile. Cela peut s'expliquer par les caractéristiques de ce dernier corpus, qui présente des textes plus longs et mieux structurés que les autres, avec des rubriques prédéfinies sur les différents aspects du jeu tels que jouabilité ou graphisme. Un autre élément intéressant de ces tests humains est d'avoir montré que, même s'ils sont légèrement supérieurs aux résultats obtenus par les logiciels des participants, les résultats des juges humains restent dans le même ordre de grandeur.

La campagne DEFT'07 a fourni des éléments d'évaluation des logiciels qui concernent une partie des problèmes soulevés en fouille d'opinion. D'autres questions importantes peuvent être évaluées : c'est le cas par exemple de la détection du caractère objectif ou subjectif global d'un texte, ou d'un segment de texte, ou bien encore des aspects multilingues de la fouille d'opinion. Si les méthodes pour traiter de la fouille d'opinion existent et fournissent des résultats honorables, les résultats obtenus par les participants à l'édition 2007 du défi prouvent que des améliorations peuvent et doivent être apportées aux modèles (avec des F-mesures classiques variant entre 0,398 et 0,566 pour le corpus le plus difficile et entre 0,457 et 0,784 pour le corpus le mieux réussi).

Parmi les améliorations à apporter aux modèles utilisés, il importe d'adapter plus précisé-

ment chaque modèle au type de corpus utilisé (domaine scientifique, style de langue, particularités lexicales, etc). Cette adaptation passe également par la recherche de solutions lors de la représentation des documents afin de mieux prendre en compte l'apprentissage et éviter les effets de sur-apprentissage et de dispersion des caractéristiques du corpus dans un nombre élevé de dimensions. En dehors de l'analyse d'opinion contenue dans des textes, de récents travaux en informatique bio-médicale sur l'extraction automatique de connaissances depuis des articles en biologie s'intéressent à la manière de déterminer la certitude et l'incertitude exprimée dans une phrase. Cette recherche s'appuie notamment sur l'étude des modalités grammaticales et des nuanceurs. À chaque phrase est ainsi accordé un indice de fiabilité relatif à l'information qu'elle contient (Jilani et Jaulent, 2009).

Les contenus porteurs d'une opinion se révèlent de plus en plus accessibles sur l'Internet, qu'ils soient collectifs et formels, par le biais de journaux en ligne, ou personnels dans la blogosphère. La fouille de textes sur ce type de documents relève d'un enjeu intellectuel et applicatif réel. Cependant, une certaine éthique reste à trouver quant à la manipulation de ces données et aux applications possibles sous-jacentes.

Références

Acosta, A. et A. Bittar (2007). La GRATOUNETTE : classification automatique générique de textes d'opinion. In *Actes de l'atelier de clôture du 3ème défi fouille de textes*, Grenoble, France.

Adda, G., J. Lecomte, J. Mariani, P. Paroubek, et M. Rajman" (1998). The GRACE French Part-of-Speech Tagging Evaluation Task. In *in Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, Volume 1, Granada, pp. 433–441. ELDA.

Adda, G., J. Mariani, P. Paroubek, M. Rajman, et J. Lecomte (1999). L'action GRACE d'évaluation de l'assignation des parties du discours pour le français. *Langues 2*(2), 119–129.

Ahat, M., W. Lenhard, H. Baier, V. Hoareau, S. Jhean-Larose, et G. Denhière (2007). Le concours DEFT'07 envisagé du point de vue de l'analyse de la sémantique latente (LSA). In *Actes de l'atelier de clôture du 3ème défi fouille de textes*, Grenoble, France.

Besançon, R., S. Chaudiron, D. Mostefa, I. Timimi, et K. Choukri (2008). The InFile project : a crosslingual filtering systems evaluation campaign. In *Proceedings of LREC 2008*, Marrakech.

Braschler, M. et C. Peters (2004). *Information Retrieval*, Volume 7, Chapter Cross-Language Evaluation forum : Objectives, results, achievements, pp. 7–31. Springer.

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

Charton, E. et R. Acuna-Agost (2007). Quel modèle pour détecter une opinion ? trois propositions pour généraliser l'extraction d'une idée dans un corpus. In *Actes de l'atelier de clôture du 3ème défi fouille de textes*, Grenoble, France.

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

Crestan, E. et R. Acuna-Agost (2007). Quel modèle pour détecter une opinion ? Trois propositions pour généraliser l'extraction d'une idée dans un corpus. In *Actes de l'atelier de clôture*

du 3ème défi fouille de textes, Grenoble, France.

Ekman, P. (1972). *Nebraska Symposium on Motivation*, Chapter Universal and cultural differences in facial expression or emotion, pp. 207–283. Lincoln : J. Cole.

Esuli, A. et F. Sebastiani (2006). SENTIWORDNET : A publicly available lexical resource for opinion mining. In *Proceedings of LREC-06, 5th Conference on Language Resources and Evaluation*, Genova, IT, pp. 417–422.

Généreux, M. et M. Santini (2007). Défi : Classification de textes français subjectifs. In *Actes de l'atelier de clôture du 3ème défi fouille de textes*, Grenoble, France.

Grouin, C., J.-B. Berthelin, S. El Ayari, T. Heitz, M. Hurault-Plantet, M. Jardino, Z. Khalis, et M. Lastes (2007). Présentation de DEFT'07 (DÉfi Fouille de Textes). In *Actes de l'atelier de clôture du 3ème DÉfi Fouille de Textes*, Grenoble, pp. 1–8. Association Française d'Intelligence Artificielle.

Habert, B., A. Nazarenko, et A. Salem (1997). *Les linguistiques de corpus*. Paris : Armand Colin/Masson.

Jilani, I. et M.-C. Jaulent (2009). Enrichissement des bases de connaissances en biologie par extraction de marqueurs de confiance dans la littérature scientifique. In M. Fieschi, P. Staccini, O. Bouhaddou, et C. Lovis (Eds.), *Risques, technologies de l'information pour les pratiques médicales*, Volume 17 of *Informatique et Santé*, pp. 113–124. Springer–Verlag France.

Mapelli, V., M. Nava, S. Surcin, D. Mostefa, et K. Choukri (2004). Technolangue : A Permanent Evaluation and Information Infrastructure. In *Proceedings of the 4th international Conference on Language Resources and Evaluation (LREC 2004)*, Lisboa, Portugal, pp. 381–384.

Maurel, S., P. Curtoni, et L. Dini (2007). Classification d'opinions par méthodes symbolique, statistique et hybride. In *Actes de l'atelier de clôture du 3ème défi fouille de textes*, Grenoble, France.

Nakache, D. et E. Métais (2005). Evaluation : nouvelle approche avec juges. In *INFORSID*, Grenoble, pp. 555–570.

Ounis, I., M. de Rijke, C. Macdonald, G. Mishne, et I. Soboroff (2006). Overview of the trec-2006 blog track. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, pp. 17–31.

Pang, B. et L. Lee (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2*(1-2), 1–135.

Peters, C. (2000). *Cross-Language Information Retrieval and Evaluation*, Volume 2069/2001 of *Lecture Notes in Computer Science*. Lisbon, Portugal : Springer.

Plantié, M., G. Dray, et M. Roche (2007). Défi DEFT07 : Comparaison d'approches pour la classification de textes d'opinion. In *Actes de l'atelier de clôture du 3ème défi fouille de textes*, Grenoble, France.

Prince, V., Y. Kodratoff, J. Azé, et M. Roche (2007). Défi Fouille de Textes : reconnaissance automatique des auteurs de discours – campagne DEFT'05 (TALN'05). *RNTI E-10*, 148 pages.

Santini, M. (2007). *Automatic Identification of Genres in Web Pages*. Ph. D. thesis, University of Brighton.

Sebastiani, F. (2005). Text categorization. In A. Zanasi (Ed.), *Text Mining and its Aplications to Intelligence, CRM and Knowledge Management*, pp. 109–129. WIT Press.

Torres-Moreno, J.-M., M. El-Bèze, F. Béchet, et N. Camelin (2007). Comment faire pour que l'opinion forgée à la sortie des urnes soit la bonne ? application au défi DEFT 2007. In *Actes de l'atelier de clôture du 3ème défi fouille de textes*, Grenoble, France.

Trinh, A.-P. (2007). Classification de texte et estimation probabiliste par machine à vecteur de support. In *Actes de l'atelier de clôture du 3ème défi fouille de textes*, Grenoble, France.

van Rijsbergen, C. (1979). *Information Retrieval*. London : Butterworths. Reprint kindly available at `http://www.dcs.gla.ac.uk/~iain/keith/`.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

Vernier, M., Y. Mathet, F. Rioult, T. Charnois, S. Ferrari, et D. Legallois (2007). Classification de textes d'opinions : une approche mixte n-grammes et sémantique. In *Actes de l'atelier de clôture du 3ème défi fouille de textes*, Grenoble, France.

Summary

From 2005 onward, the French DEFT national evaluation campaigns have been offering exploratory topics in text mining. The 2007 challenge was about classifying opinion texts: the task consisted in assigning an opinion class to each text in a corpus, among 2 or 3 classes from an unfavorable to a favorable judgment. Four corpora were made available to the participants: parliamentary debates over a draft law, comments about video games, reviews of books and films, and reviews of scientific papers. In this paper, we first describe the preparative stage of the campaign, including corpus collection, definition of evaluative measurements, and human tests of the task. In a second part, we present an analytic overview of results obtained by the participants, along with subsequent remarks about the different kinds of corpora. Finally, we develop a synthetic assessment of methods that were submitted to evaluation.