DEFT2014, analyse automatique de textes littéraires et scientifiques en langue française

Charlotte Lecluze Gaël Lejeune
Université de Caen
GREYC, CNRS, CS14032, 14032 Caen Cedex 5
prenom.nom@unicaen.fr

Résumé. Nous présentons dans cet article les méthodes utilisées par l'équipe HULTECH pour sa participation au DÉfi Fouille de Textes 2014 (DEFT2014). Cette dixième édition comportait quatre tâches et portait sur l'analyse automatique de textes littéraires et d'articles scientifiques en langue française. Les trois tâches portant sur l'analyse de textes littéraires consistent à évaluer le genre d'une part mais aussi la qualité littéraires des nouvelles mises à notre disposition. La dernière tâche quant à elle porte sur l'analyse de textes scientifiques, à savoir des articles des sessions précédentes de TALN. Notre équipe a participé aux quatre tâches.

Abstract. **DEFT2014, automatic analysis of literary and scientific texts in French**
We present here the HULTECH (HUman Language TECHnology) team approach for the DEFT2014 (french text mining challenge). The purpose of these four tasks challenge is to automatically analyze a special kind of literay texts : short stories. The last one is about scientific articles. The three tasks about short stories aim to detect the genre, to assess the quality of the text and the consensus between reviewers about this quality. The last task relates to the analysis of scientific texts: articles of previous sessions of TALN. Our team participated in all of the four tasks.

Mots-clés : classification, évaluation, algorithmique du texte, stylométrie.

Keywords: classification, evaluation, text algorithmics, stylometry.

1 Introduction

Pour cette nouvelle édition du défi, quatre tâches d'analyse de texte étaient proposées. Les trois premières tâches s'attachaient aux textes littéraires (courtes nouvelles) tandis que la quatrième concernait des articles scientifiques :

Tâche 1 - classification par genre littéraire Catégoriser le genre littéraire de courtes nouvelles parmi 45 sous-catégories (poésie, nouvelles, policier...Voir tableau 1). Cette tâche consistait à classer automatiquement une nouvelle dans la sous-catégorie qui lui a été la plus souvent attribuée par des annotateurs. Chaque nouvelle avait été classée dans 2 ou 3 sous-catégories différentes par les annotateurs. Les sous-catégories d'une œuvre sont ordonnées par ordre d'importance : la première est la sous-catégorie principale...(Figure 1 et 2).

Tâche 2 - évaluer la qualité littéraire Évaluer la qualité littéraire de chacune de ces nouvelles en prédisant la note que donnerait un juge humain. La tâche 2 a pour but d'évaluer la qualité littéraire de chacun de ces textes en prédisant la note attribuée par le comité de relecture à chacun des textes littéraires. La référence de cette tâche est constituée par l'ensemble des notes attribuées par le comité de relecture de l'éditeur Short Edition. Ces notes ont été fournies avec le corpus d'entraînement. Il s'agit pour chaque nouvelle d'une série de 3 à 13 annotations auxquelles une note de 1 à 6 a été associée. L'analyse des commentaires des relecteurs doit permettre l'évaluation de la qualité littéraire de la nouvelle et l'attribution automatique d'une note.

Tâche 3 - évaluer le consensus sur la qualité Déterminer, pour chacune des nouvelles, si elle est consensuelle auprès des différents relecteurs. Une œuvre est jugée consensuelle si les notes attribuées par les différents relecteurs ne varient pas au-delà d'un écart de 1 point.

Tâche 4 - classer par session scientifique Cette tâche se démarque des précédentes car elle concerne les articles scientifiques présentés lors des dernières conférences TALN. Pour chaque édition précédente de TALN, identifier dans quelle session scientifique chaque article scientifique de la conférence a été présenté (communication orale uniquement), parmi la liste fournie pour chaque édition. Pour chaque édition, un ensemble d'articles (titre, résumé, mots-clés, texte), la liste des sessions scientifiques de cette édition, et la correspondance article/session (sauf pour le test) ont été fournis. Le corpus de test se composait quant à lui d'une édition complète de TALN (articles et liste des sessions) pour laquelle il fallait identifier dans quelle session chaque article a été présenté.

```
<post>
<id>1575</id>
<title><![CDATA[Tant de temps]]></title>
<content><![CDATA[<p>Il y a un temps pour tout<br />
Un temps atout<br />
Quitte à tout prendre<br />
Je prends le temps.</p>
]]></content>
<type><![CDATA[poetik]]></type>
<subcategories>
<subcategory>
<id>1421</id>
<section>poésie</section>
<name>haikus</name>
<rank>0</rank>
</subcategory>
<subcategory>
<id>1443</id>
<section>émotions</section>
<name>instant de vie</name>
<rank>1</rank>
</subcategory>
</subcategories>
</post>
```

FIGURE 1 – Exemple de nouvelle du corpus d'apprentissage de la tâche 1 : contenu textuel et méta-informations

2 Description des tâches

2.1 Tâche 1 : classification de courtes nouvelles

L'objectif était de proposer un classement des nouvelles par genre littéraire. 45 sous-catégories, réparties en sept section, étaient envisagées. Le tableau 1 présente l'effectif des nouvelles par section et sous-catégorie dans le corpus d'apprentissage mis à notre disposition et qui contenait 2328 nouvelles, chacune s'étant vu attribuer deux à trois sous-catégories (Figure 1 et 2) par les relecteurs, pour un total de 5182 annotations. Ce tableau illustre la disproportion de certaines sections par rapport à d'autres. Les sections *chronique*, *émotions* et *poésie* sont notamment surreprésentées.

La figure 1 présente l'exemple d'une nouvelle classée par les annotateurs dans deux sections - sous-catégorie, principalement dans la section - sous-catégorie *Poésie* - *haikus* mais aussi dans celle *Émotions* - *Instant de vie* (en bleu sur la figure).

Les réponses à prédire pour l'ensemble des nouvelles ont été fournies dans un fichier tabulaire (4 colonnes : numéro de document, sous-catégorie, section, rang), comme le montre la figure 2. Dans une phase d'exploration du corpus d'apprentissage mis à notre disposition, nous avons mesuré d'une part les combinaisons de sous-catégories qu'il était possible de rencontrer (Tableau 2), ainsi que les catégories les plus fréquemment attribuées au rang 1 (Tableau 3).

Section	Effectifs	Sous-catégories	Sous-effectifs
autres	256	5	[allégorie:39,autres:6, conte:59, fantaisiste:143, lettre:9]
chronique	1166	9	[arts:200, gastronomie:14, histoire:91, nature:268, sciences-médical:33, société:440, spiritualité:40, sport:16, voyage:64]
émotions	2614	11	[amitié:80, colère:39, drame:355, enfance:131, erotisme:46, famille:290, humour:196, instant de vie:586, mélancolie-solitude:279, nostalgie-souvenirs:255, romance:357]
fantastique-sf	146	5	[esotérique:3, fantasy:4, merveilleux:40, science-fiction:45, surnaturel:54]
jeunesse	29	1	[jeunesse:29]
noir	175	5	[aventure:25, horreur:14, policier:23, suspens:96, thriller:17]
poésie	796	9	[alexandrins:140, chanson:26, comptine:20, fable:13, haikus:3, prose:32, slam:14, sonnets:40, vers libres:508]
Total	5182	45	

Tableau 1 – Effectifs des section et sous-catégories du corpus d'apprentissage de la tâche 1

1575	haikus	poésie	0
1575	instant de vie	émotions	1

FIGURE 2 – Extrait du fichier des réponses à prédire

Combinaisons de sous-catégories	Effectif
Chronique - société / Émotions - instants de vie	120
Émotions - famille / Émotions - drame	101
Émotions - romances / Poésie - Vers libres	87
Émotions - romances / Émotions - nostalgie-souvenirs	51
Fantastique-sf - merveilleux / Chronique - nature	7
Poésie - haikus / Émotions - instants de vie	1
Poésie - haikus / Nostalgie - souvenir	1

Tableau 2 – Exemples de combinaisons de sous-catégories rencontrées avec leur effectif

Nous avons également mesuré les étiquettes les plus souvent attribuées. Le tableau 3 présente les effectifs des dix sous-catégories principalement attribuées.

Effectif	Sous-catégories
496	Poésie - vers libres
246	Émotions - instant de vie
190	Chronique - société
152	Émotions - drame
140	Poésie - alexandrins
111	Émotions - famille
86	Chronique - arts
83	Chronique - nature
75	Émotions - romance
63	Autres - fantaisiste

Tableau 3 – Les dix étiquettes au rang 1 les plus souvent attribuées par les annotateurs.

20914	1	3.0
20914	2	3.0
20914	3	2.0

FIGURE 3 – Extrait du fichier résumant les notes attribuées

<post>
<id>20914</id>
<title><![CDATA[Les hérons usés]]></title>
<content><![CDATA[Comme vieux flamants se querellent, et disparaissent les rebelles
 Comme nos aigles tatoués aussi vont s'envoler
 Comme demain nous serons résignés
 Comme des hérons usés

 Comme ces oiseaux posés sur des arbres âgés
 Et l'étang millénaire est d'essence et d'éther
 Comme trop de peine, trop peu d'oxygène
 Comme leurs forêts s'éteignent

 Comme ces hérons parés, force grêle et cendrée
 Comme ils regardent au loin le château des humains
 Comme leur monde souffre et saigne
 Ils sont heureux quand même
 Heureux quand même]]></content>
<type><![CDATA[poetik]]></type>
<reviews>
<review>
<id>1</id>
<uid>56766</uid>
<content><![CDATA[
rhéron, héron, petit patapon...
Un petit oui aussi.
]]></content>
<note>
3.0
</note>
</review>
<review>
<id>2</id>
<uid>28729</uid>
<content><![CDATA[
J'aime bien cette déclinaison dans le rythme des strophes mis à part cela c'est très léger
]]></content>
<note>
3.0
</note>
</review>
<review>
<id>3</id>
<uid>8519</uid>
<content><![CDATA[
Un joli sujet, une mélancolie communicative.
]]></content>
<note>
2.0
</note>
</review>
</reviews>
</post>

FIGURE 4 – Exemple de nouvelle du corpus d'apprentissage de la tâche 2 : contenu textuel et métadonnées (dont les notes attribuées par les annotateurs)

<post>
<id>5084</id>
<title><![CDATA[Et le vent souffla: Comme un carré de liège]]></title>
<content><![CDATA[Tout en caracolant
 Le sommeil me ravit
 A la prison de mon lit
 Je bondis, léger,léger
 Comme un carré de liège

 ]]></content>
<type><![CDATA[tres-tres-court]]></type>
<reviews>
<review>
<id>1</id>
<uid>24799</uid>
<content><![CDATA[
Poétik, non ? Un peu court, un peu faible.
]]></content>
<note>
4.0
</note>
</review>
<review>
<id>2</id>
<uid>27092</uid>
<content><![CDATA[
Improvisation?
]]></content>
<note>
4.0
</note>
</review>
<review>
<id>3</id>
<uid>10612</uid>
<content><![CDATA[
TTC ou poétik ? Ni l'un ni l'autre. C'est faible.
]]></content>
<note>
5.0
</note>
</review>
</reviews>
<consensus>
<decision>1</decision>
</consensus>
</post>

FIGURE 5 – Exemple de nouvelle du corpus d'apprentissage de la tâche 3 : contenu textuel et métadonnées dont les notes attribuées par les annotateurs et la décision quant au consensus

2.2 Tâche 2 : évaluation de la qualité littéraire

Á partir de l'analyse des commentaires réalisés par les annotateurs, la tâche 2 consistait à estimer la qualité littéraire de chaque nouvelle. Chaque nouvelle bénéficiait de trois à treize annotations, consistant en quelques mots (en bleu sur la figure). Chaque annotation était assortie d'une note entre un et six (en rouge sur la figure). La figure 4 présente une des nouvelles du corpus d'apprentissage de la tâche 2. Sur cette nouvelle, trois annotateurs ont donné leur avis. « Id » est le numéro de la relecture pour ce document, « uid » est l'identifiant numérique du relecteur (un seul identifiant par relecteur sur l'ensemble du corpus), « content » correspond au contenu de la relecture, « note » correspond à la note attribuée par le relecteur.

Un fichier (Figure 3) contenant un résumé des notes attribuées par les relecteurs était fourni avec le corpus d'apprentisssage. Ce fichier comprenait sur une ligne l'identifiant de la nouvelle, le numéro de la relecture pour ce document et la note attribuée par le relecteur.

2.3 Tâche 3 : déterminer si une œuvre fait consensus

La tâche 3 consistait à évaluer si l'œuvre à analyser faisait consensus ou pas, autrement dit si les notes attribuées par les différents relecteurs ne varient pas au-delà d'un écart de 1 point.
Avec le corpus d'apprentissage, un fichier mentionnant pour chaque nouvelle si elle faisait consensus ou pas (avec 0 = œuvre non consuelle et 1= œuvre consensuelle) était fourni. Chaque ligne du fichier comprend l'identifiant de nouvelle et la décision.

2.4 Tâche 4 : déterminer la session scientifique

Cette tâche consistait à affecter automatiquement un article scientifique à une session scientifique thématique. En l'occurrence, les participants avaient accès au nom de la session. Ce nom pouvait dans certains cas être composé de plusieurs mots ou de thèmes différents. Dans ce dernier cas, les différents thèmes étaient séparés par des « | ». Les articles avaient été extraits automatiquement à partir des fichiers PDF. certaines marques telles que les légendes, les titres ou les notes de bas de page se sont trouvées « écrasées »par ce processus. Ces problèmes d'écrasement des observables sont classiques lors du passage automatique du PDF au XML ou HTML.

3 Méthodologie et résultats

3.1 Tâches 1 et 3 : stylométrie

Notre hypothèse était la suivante : la manière la plus économe de rattacher un texte à une catégorie est de se fier à des indices stylométriques. Ainsi, nous avons considéré ces deux tâches comme similaires à la désanonymisation d'articles (ou *Autorship Attribution*). L'objet étant ici de détecter un style collectif plutôt qu'un style individuel. Ce style collectif serait spécifique à un sous-genre (tâche 1) ou pourrait amener un consensus auprès des relecteurs (tâche 3). Nous y avons ajouté des critères grammaticaux avec l'utilisation des pronoms personnels et des auxiliaires. Il nous a semblé que des indices purement lexicaux seraient moins robustes, néanmoins nous avons proposé d'ajouter quelques champs lexicaux bien déterminés.

Une hypothèse envisagée mais non retenue avait été de chercher des patrons linguistiques dans l'esprit des travaux de Bechet *et al.* (2012). On pourrait en effet penser que certaines structures phrastiques ou sous-phrastiques soient typiques de certaines classes.

3.1.1 Critères utilisés

Les critères stylométriques (Stamatatos, 2009; Sun *et al.*, 2012) retenus font partie des critères classiquement utilisés dans le domaine.

Nous avons testé l'ajout de plusieurs champs lexicaux. Seul celui que nous avons nommé « saisons » apportait une plus-value, principalement sur la classification des œuvres.

Ce champ lexical, construit manuellement, comportait les noms de saisons ainsi que quelques termes connexes : fleur(s), feuille(s), neige, soleil…

3.1.2 Spécificités de chaque *run* et résultats

Pour ces deux tâches, nous avons testé différents algorithmes d'apprentissage pour construire nos modèles. La configuration la plus robuste dans notre cas était un classifieur SVM pour lequel la phase d'apprentissage était effectuée à l'aide

Critères stylométriques		Critères grammaticaux
Taille :	# Paragraphes	#Première personne singulier
	# Phrases	#Deuxième personne singulier/pluriel
	# Mots	#Troisième personne singulier
	# Caractères	#Autres personnes (nous, ils, eux)
Ponctuation :	# Virgules	#Pronoms relatifs
	# Point-virgules	#Occurrences être/avoir au présent
	# Tirets	#Occurrences être/avoir au passé
	# Parenthèses	#Occurrences être/avoir au futur
	# Guillemets	#Termes comparaison métaphores (comme, tel...)

Tableau 4 – Critères stylométriques et grammaticaux utilisés

de l'algorithme SMO (Platt, 1999). Le jeu sur les paramètres n'a eu qu'un impact marginal sur les résultats. Pour tous les *run*s présentés, nous avons donc conservé les valeurs par défaut de l'implémentation présente dans *Weka*.

Pour la tâche 1 où la sortie était une liste ordonnée de classes, nous utilisions la méthode suivante :
– le classifieur nous donne l'étiquette la plus prégnante que nous plaçons au rang 1 ;
– les étiquettes suivantes sont déduites des associations les plus probables observées dans le corpus d'apprentissage (Tableau 2).

run	Spécificités	Résultat	Rang
Tâche 1, *run* 1	Critères stylométriques et grammaticaux	NDCG : 0,513	N/A
Tâche 1, *run* 2	Ajout du champ lexical « saisons »	NDCG : 0,5248	1er/3
Tâche 3, *run* 1	Critères stylométriques et grammaticaux	Précision : 0,3776 (soumis) 0,5449 (réel) [1]	2ème/2

Tableau 5 – Spécificité de chaque *run* et résultats

3.2 Tâche 2 : motifs récurrents

3.2.1 Critères utilisés

Pour cette tâche, nous nous sommes intéressés à des motifs répétés trouvés dans les critiques. Ces motifs sont des N-grammes de caractères avec $n > 3$ de manière à éviter la surgénération de motifs. En effet, ces motifs courts étaient peu discriminants et induisaient une charge de calcul importante. Nous ne conservions que les motifs répétés dans le corpus, c'est-à-dire ceux qui étaient présents dans au moins deux critiques. Si nous reprenons la terminologie de la fouille de texte, nous avons donc des motifs avec un support (dans les critiques) strictement supérieur à 1 et une longueur strictement supérieure à 3. Nous avons ainsi obtenu 1322 motifs, quelques exemples sont donnés le tableau 6.

Dommage	Trop_de_fautes	J'aime_beaucoup	histoire_m	que_j_ai_lu
Encore	Un_peu	J'aime_l	humour	sans_faute
Ennuyeux	Aucun	Je_suis	intéressante	extraordinaire
Faible	Aucun_intérêt	Je_vais	je_me	pas_terrible
Je_n'ai	Beaucoup	l'auteur	jusqu'à_la_fin	Un_petit_oui

Tableau 6 – Exemples de motifs extraits (« _ » représentant une espace typographique)

Pour construire le modèle nous avons utilisé la même configuration que pour les tâches 1 et 3.

run	Spécificités	Résultat	Rang
Tâche 2, *run* 1	Motifs répétés, longueur supérieure à 3 caractères	EDRM : 0,3975	2ème/2

Tableau 7 – Description synthétique de notre *run* de la tâche 2

3.3 Tâche 4 : une *baseline*?

3.3.1 Critères utilisés

Étant donné les noms des sessions disponibles, nous avons supposé que ces termes seraient présents dans les textes et que le nom de la session correspondante serait plus fréquent que les autres.

Pour le *run* 1 nous avons simplement cherché l'effectif de chaque terme dans l'article.

Dans un second temps, nous avons ajoutés des critères positionnels. Ces termes sont d'autant plus importants qu'ils sont placés en tête (introduction/première section) et pied de document (conclusion/bibliographie).

Pour le *run* 2 nous avons donc donc retiré le corps du document de manière à ne considérer que les occurrences placées en tête et en pied de document. Pour le *run* 3, nous avons cherché à affiner cette hypothèse. Les termes ont d'autant plus de poids que leurs occurrences sont :
– Proches de la tête ou du pied (donc distants du milieu du document)
– Présentes dans des petits segments de textes (*a priori* : mot-clés, titres, légendes, entrées bibliographiques)
Pour le *run* 3, le calcul du poids d'un terme est la moyenne du poids de chacune de ses occurrences dans l'article. Le poids de chaque occurrence est calculé comme suit :

$$\frac{DistCenter}{len(article)} + \frac{1}{len(segment)}$$

Avec $len(X)$, une fonction donnant la longueur de X en caractères et $Distcenter$ la distance (en caractères) entre la position de l'occurrence et le centre du document ($\frac{len(document)}{2}$).

run	Spécificités	Résultat	Rang
Tâche 4, *run* 1	Effectif des noms de session, article complet	Précision au rang 1 : 0,4259	N/A
Tâche 4, *run* 2	Effectif des noms de session, article évidé	Précision au rang 1 : 0,4814	3ème/5
Tâche 4, *run* 3	Effectif pondéré des noms de session	Précision au rang 1 : 0,4444	N/A

Tableau 8 – Descriptions synthétiques des *run*s pour la tâche 4

4 Conclusion

En conclusion de notre participation au DEFT 2014, nous souhaitons souligner qu'une fois de plus les organisateurs ont fait preuve d'originalité avec une édition orientée autour d'un genre rarement explorée : la nouvelle. Les 3 tâches offraient des perspectives de recherches intéressantes même si la tâche 2 (évaluer la qualité littéraire) aurait pu proposer une piste basée uniquement sur les textes eux mêmes. Pour ce qui est de la tâche 4, une idée qui nous paraît intéressante serait de proposer un travail directement sur les fichiers PDF. En effet, la phase de conversion amène des déperditions d'informations et globalement l'écrasement de certains observables (les tableaux par exemple). Selon la méthode de traitement que l'on souhaite appliquer en aval, ces pertes seront plus ou moins handicapantes. Proposer les documents « bruts » permettrait d'évaluer, de manière indirecte, l'influence des pré-traitements sur les résultats. En effet, ces problèmes ne relèvent pas que de l'ingénierie et sont de vrais objets de recherche. On pourrait même imaginer que le travail des participants soit justement de concevoir un système de pré-traitement. Ce système serait évalué en fonction des résultats d'un autre module de traitement placé en aval. Le système de pré-traitement serait dès lors évalué non seulement de façon intrinsèque (quantité de mots, de phrases conservées...) mais aussi de manière extrinsèque en fonction de son influence sur les résultats d'un module de post-traitement.

Références

NICOLAS BÉCHET, PEGGY CELLIER T. C. & CRÉMILLEUX B. (2012). Discovering linguistic patterns using sequence mining. In *Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2012)*, p. 11–17.

PLATT J. C. (1999). Advances in kernel methods. chapter Fast Training of Support Vector Machines Using Sequential Minimal Optimization, p. 185–208. Cambridge, MA, USA: MIT Press.

STAMATATOS E. (2009). A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, **60**(3), 538–556.

SUN J., YANG Z., LIU S. & WANG P. (2012). Applying stylometric analysis techniques to counter anonymity in cyberspace. *Journal of Networks*, **7**(2).