Master 2 ATAL

Corpus et Méthodes Expérimentales

HTML Killer

Étudiants :
Marie-Charlotte Daureu,
Coraline Marie et
Carl Goubeau

Encadrant :
Gael Lejeune

6 janvier 2015

UNIVERSITÉ DE NANTES

1 Contexte

Les corpus sont de précieux outils absolument indispensables pour les scientifiques du Traitement Automatique des Langues. En effet, le corpus sert à la fois de support de travail, de motivation, de justification, et de modèle pour bien des domaines en rapport avec le TAL. Ils peuvent être immenses et contenir des millions de données. Il peuvent également exister sous de multiples formes (textuelle, sonore, ...). Malheureusement, les corpus sont rares, chers et très difficiles à construire de façon propre et convenant à de multiples usages.

Aujourd'hui, la création de corpus est devenu un véritable challenge pour les scientifiques du TAL. Utiliser un corpus est facile, mais en créer un qui correspond parfaitement à nos besoins est beaucoup moins aisé. Il existe même des concours afin de motiver les chercheurs à trouver de nouveaux moyens de créer des corpus. L'une des méthodes les plus naïves permettant de collecter un grand nombre de données textuelles, consiste à simplement récupérer du texte sur des pages web. Malheureusement ces textes sont extrèmement bruités du fait que l'HTML n'est pas du tout un langage strict.

Dans le cadre du cours *Corpus et Méthodes Expérimentales*, dispensé à l'Université de Nantes, nous devions étudier différents outils spécialisés dans le nettoyage des pages web. De plus nous devions également coder un programme simpliste, utilisant des méthodes naïves pour nettoyer des page web. Ces différents outils ont été testés sur des corpus contenant des pages web de cinq langues différentes. Dans ce rapport nous présenterons donc les corpus qui nous ont servi pour les tests. Nous présenterons ensuite les différents outils utilisés ainsi que notre baseline, avant de terminer par la présentation des résultats que nous avons obtenus.

2 Présentation des outils

2.1 Baseline

Nous avons donc développé un script permettant de nettoyer les fichiers sources d'un site internet. Ce dernier peut, au choix, produire le resultat en texte brut, ou une nouvelle page web contenant les balises HTML suivantes : <p> et <h>.

Après lecture de plusieurs fichiers sources, nous avons choisi de ne pas conserver le code se trouvant avant le titre principal de la page, c'est-à-dire la première balise <h1>. Nous ignorons de la même manière ce qui se trouve après la balise fermante du corps de la page </body>. Cette heuristique permet de supprimer une grande partie de code CSS et javascript.

Nous recherchons dans les pages les balises suivantes afin de les supprimer, ainsi que leurs contenus : <script>, <img>, <!– commentaires –>, <form> et <li>.

D'autres balises possèdent quant à elles du contenu interessant, nous ne

supprimons donc que les balises, et non le texte qu'elles contiennent : les liens <a>, et les .

2.1.1 Les titres <h>

Pour retrouver le titre principal de la page traitée, nous cherchons dans le code le contenu de la première balise <h1>. Les autres élements HTML de hierarchisation sont supprimés (h2, h3, ...).

2.1.2 Les paragraphes

Les pages web ne respectant pas une structure unique, cette partie fût plus délicate à réaliser. En effet, certaines pages possèdent leurs contenus entre des balises de paragraphe , d'autres entre , et certaines à même le corps de la page : <body> ...

Nous nous sommes alors appuyés sur les retours à la ligne
 pour segmenter le fichier en paragraphes. Afin de prendre en compte les structures HTML, pour les sites qui l'utilisent, nous remplaçons au préalable les fins de paragraphes par des balises
.

2.2 BoilerPipe

La bibliothèque BoilerPipe [1] fournit des algorithmes pour détecter et supprimer le surplus non pertinent autour du contenu principal d'une page web. Cet outil ne nécessite pas de données supplémentaires pour fonctionner. Il est néanmoins possible de passer en paramètre la strategie d'extraction pour une meilleure efficacité. BoilerPipe propose en sortie d'obtenir soit des fichiers avec des balise HTML, ou bien du texte pur. BoilerPipe utilise un algorithme d'apprentissage automatique nettoyer les pages.

2.3 Readability

Readability[2] est un outil ayant pour but de rendre meilleure l'expérience de lecture sur les sites internet. Cet outil permet à l'utilisateur de personnaliser l'affichage du site visité, en ne conservant que les informations pertinentes. De nos jours, il est utilisé par un grand nombre d'internautes pour tranformer le Web en un endroit plus agréable à lire.

Le base de Readability se sert de règles pour nettoyer les corpus. Cet outil est utilisé par de nombreuses applications, telles que Safari, Kindle (Amazon), Flipboard, ...

Nous utilisons ici le noyau de Readability qui nous permettera de nettoyer les pages de notre corpus.

3 Présentation des corpus

Pour tester et évaluer l'ensemble des outils que nous avons présenté, nous avons utilisé un corpus multilingue de pages web sur le thème de la santé. Il

contient des fichiers dans cinq langues différentes : grecque, anglaise, polonaise, russe et chinoise. Il y a donc un corpus brut, qu'il faut nettoyer et un corpus gold, qui contient les fichiers de réference, que l'on souhaite obtenir. Les données ont été récupérées entre 2011 et 2012, et du fait des langues utilisées, contient cinq alphabets différents.

Au terme de ce travail, nous obtenons trois corpus supplémentaires, qui contiennent les fichiers nettoyés par les différents outils que nous avons étudiés.

Avant de pouvoir évaluer et comparer les outils que nous avons utilisés, nous avons commencé par faire quelques recherches sur les corpus. Pour cela, nous avons créé un programme permettant de relever quelques statistiques sur les corpus.

Les fichiers HTML sont triés par langue. On remarque ainsi qu'il y a deux fois plus de fichiers en anglais ou en chinois que dans les autres langues. Le nombre de mots en chinois est considérablement inférieur à celui des autres langues, car la ligne de commande utilisée pour produire les statistiques ne peut fonctionner correctement sur une écriture non segmentée (en mots) telle que le chinois ou le japonais.

Langue du corpus	Nombre de paragraphes	Nombre de lignes	Nombre de mots	Nombre de caractères	Nombre de fichiers
grec (el)	866	516581	2172414	31783683	273
anglais (en)	7875	823076	2363492	38811928	475
polonais (pl)	4725	279697	1129605	18176086	274
russe (ru)	3901	323108	1195397	21818585	267
chinois (zh)	4652	388332	1101700	25116896	405

Table 1 – Statistiques pour le corpus brut

Langue du corpus	Nombre de paragraphes	Nombre de lignes	Nombre de mots	Nombre de caractères	Nombre de fichiers
grec (el)	2659	4144	116398	1426722	273
anglais (en)	7269	9220	225764	1453612	475
polonais (pl)	3084	4812	115872	900604	274
russe (ru)	2035	4393	70331	936467	267
chinois (zh)	4409	5310	12278	1133436	405

Table 2 – Statistiques pour le corpus gold

Au terme de ces divers statistiques, on peut remarquer que Boilerpipe ne conserve pas les balises de paragraphe, contrairement à notre baseline, au gold et à Readability, il y a donc de la perte d'information concernant la structure du document. On remarque également que notre baseline est beaucoup plus stricte dans son nettoyage que les autres outils. Cependant, ce n'est pas une

Langue du corpus	Nombre de paragraphes	Nombre de lignes	Nombre de mots	Nombre de caractères	Nombre de fichiers
grec (el)	687	1794	26503	464836	273
anglais (en)	7514	9810	198690	1271535	475
polonais (pl)	4372	6217	122633	919232	274
russe (ru)	2366	3729	67194	910438	267
chinois (zh)	2665	4491	13172	683570	405

TABLE 3 – Statistiques pour le corpus nettoyé par Html_killer

Langue du corpus	Nombre de paragraphes	Nombre de lignes	Nombre de mots	Nombre de caractères	Nombre de fichiers
grec (el)	0	5851	131153	1742018	273
anglais (en)	0	6546	259492	2080506	475
polonais (pl)	0	4134	126863	1155244	274
russe (ru)	0	3278	105790	1649710	267
chinois (zh)	0	3650	20941	1136238	405

TABLE 4 – Statistiques pour le corpus nettoyé par Boilerpipe

Langue du corpus	Nombre de paragraphes	Nombre de lignes	Nombre de mots	Nombre de caractères	Nombre de fichiers
grec (el)	1006	13629	132448	1760130	273
anglais (en)	4252	18696	249166	1222357	475
polonais (pl)	1736	10440	121145	1114513	274
russe (ru)	1442	10870	80059	1265858	267
chinois (zh)	2415	11922	18919	1260021	405

TABLE 5 – Statistiques pour le corpus nettoyé par Readability

bonne chose car cela signifie qu'il y a de la perte d'information au niveau du contenu de la page. Readability semble conserver beaucoup plus d'informations au niveau du nombre de lignes mais pas au niveau du nombre de caractères. Par conséquent, la comparaison entre ces outils sera plus significative avec l'aide de CleanEval.

4 Évaluation

4.1 CleanEval

Cleaneval[3] est une évaluation compétitive et partagée sur le nettoyage de pages web, ayant pour but de préparer les données récupérées sur internet pour qu'elles puissent être utilisées en corpus pour de la linguistique et du développement de technologies du langage. Un script d'évaluation de Cleaneval nous permet de calculer les F-mesure, Précision et Rappel des différents outils pré-

sentés dans ce rapport.

4.2 Résultats

Langue	F-mesure	Précision	Rappel
Grec	17.27	46.29	10.61
Anglais	61.68	66.13	57.78
Polonais	57.58	56.16	59.06
Russe	45.45	46.51	44.44
Chinois	**46.73**	47.51	45.98

TABLE 6 – Résultats de la Baseline pour le nettoyage de pages web

Langue	F-mesure	Précision	Rappel
Grec	**84.34**	79.60	89.68
Anglais	77.47	71.99	83.84
Polonais	74.37	71.39	77.62
Russe	56.52	47.07	70.74
Chinois	8.06	6.54	10.49

TABLE 7 – Résultats de BoilerPipe pour le nettoyage de pages web

Langue	F-mesure	Précision	Rappel
Grec	83.56	78.37	89.48
Anglais	**83.92**	79.72	88.59
Polonais	**76.83**	74.92	78.84
Russe	**69.72**	65.10	75.05
Chinois	31.41	26.27	39.04

TABLE 8 – Résultats de Readability pour le nettoyage de pages web

4.3 Discussions

Grâce à ces résultats, nous pouvons remarquer que Readability semble être le meilleur outil que nous ayons testé, pour nos besoins. En effet, sur les tables précédentes, Readability obtient les meilleurs scores de F-mesure pour l'Anglais, le Polonais et le Russe. Boilerpipe n'est pas très loin et obtient les meilleurs résultats pour le Grec. Ce qui est surprenant, c'est que notre baseline possède le meilleur score de F-mesure en ce qui concerne le Chinois.

Au terme de ce travail, plusieurs constatations peuvent être faites. Tout d'abord, nous avons remarqué que le nettoyage de pages Web n'est pas encore évident. Il est très difficile de nettoyer correctement de nombreux fichiers HTML

ayant des formes et des provenances différentes. Ceci peut s'expliquer pour les raisons suivantes :

La diversité des langues provoque des problèmes d'encodage. Ceci se remarque surtout avec les pages web en Chinois, car BoilerPipe ne peut pas reconnaître tous les caractères, et les interprêter.

Le HTML n'est pas un langage de programmation strict. Ainsi, la majorité des fichiers HTML comporte des erreurs dans leur programmation, et bien qu'elles n'influent pas sur l'aspect visuel de la page web fourni par un navigateur, elles sont responsables d'une grande partie des erreurs de nettoyage (balises manquantes, non respect des structures, ...).

Les pages web peuvent avoir de multiples formes (forums, journaux, blogs, ...). Il est donc également très difficile pour les outils de nettoyage de s'adapter à cette diversité qui ne cesse d'augmenter.

Le dernier problème que l'on peut signaler est celui des conventions de nettoyage. En effet, bien que nous ayons travaillé avec un corpus gold pour l'évaluation, les outils de nettoyage tel que Boilerpipe ou Readability ne sont pas forcément créés pour s'adapter à ce même standard. Nous l'avons remarqué surtout avec BoilerPipe, car il ne conserve pas les balises de paragraphe, alors que le gold et Readability les conservent.

5 Conclusion

En tant qu'étudiants en Traitement Automatique de Langues, nous avons prit l'habitude de travailler sur des corpus, mais moins de les créer nous même. Ce projet nous a donc permis de nous sensibiliser et de nous faire prendre conscience de l'importance et de la difficulté de mettre au point des corpus robustes et pertinents à partir de sources libres tel que le web.

Bien qu'aujourd'hui, les technologies s'appuyant sur le TAL soient de plus en plus nombreuses, elles ne sont pas encore optimales, et il reste beaucoup à faire. Cependant, de bonnes choses existent déjà, et nous pouvons aisément espérer que la création de corpus, ou le nettoyage de fichiers HTML deviennent des tâche facilement réalisable dans un futur proche.

En ce qui concerne le code et les résultats créés au cours de ce projet, ils sont disponibles sur un dépôt GitHub : https://github.com/Slayerxoxo/CME_Project.

Références

[1] BoilerPipe. http://code.google.com/p/boilerpipe/. [Ressource en ligne disponible au 06 Janvier 2015].

[2] Readability. https://readability.com. [Ressource en ligne disponible au 06 Janvier 2015].

[3] Adam Kilgarriff Marco Baroni, Francis Chantree and Serge Sharoff. Cleaneval : a competition for cleaning web pages. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http ://www.lrec-conf.org/proceedings/lrec2008/.