Atelier d'écriture scientifique

VGT9E52

Support de cours : exercices

NICOLAS GUTEHRLÉ

Réécriture de texte

Les textes ci-dessous sont des extraits de mémoires de recherches ayant des défauts rédactionnels variés. Réécrivez ces textes en appliquant les conseils vus dans le cours :

Ci-dessous le calcul de la distance entre « avion » /avjɔ̃/ et « camion » /kamjɔ̃/ à titre d'exemple

Avion	Camion	Distance	Distance totale
/a/	/k/	6	
/v/	/a/	6	
/j/	/m/	6	
/ɔ̃/	/j/	5	
X	/ɔ̃/	1,5	24,5

Tableau 1 : Calcul de la distance entre les voyelles et consonnes des mot « avion » et « camion »

Comme indiqué dans l'ouvrage de Monique et Pierre Léon paru en 2007, les cas d'assimilations consonantiques peuvent être du à des facteurs physiques ou psychologiques.

Nous avons extrait les 9 règles ci-dessous après avoir contrôlé manuellement les phrases obtenues après avoir substitué un mot par un de ses mots proches ou bien en contrôlant un mot dans un contexte à partir des 50 phrases que nous avons constitué.

Notre méthode de travail consistera à analyser les articles sur l'interface web de l'outil ALP, ce dernier est un outil gratuit et accessible sur le site internet *arabicnlp.pro/alp*.

Ils possédaient cependant de nombreuses restrictions sur la manière d'utiliser la langue, mettant ainsi en question leur qualification de SQR puisque les requêtes doivent être formulées en langue naturelle.

Un problème majeur a également été rencontré : celui de la taille du corpus. Dans le corpus d'entraînement, certaines classes étaient fortement sous-représentées par rapport à d'autres, ce qui a faussé leur classification.

Ce modèle de détection de la similarité est selon nous pertinente dans le sens où les questions posées par des utilisateurs dans des FAQ ou des forums sont rarement directes, peuvent être grammaticalement ou orthographiquement incorrectes et peuvent prendre plusieurs tournures différents (« est-ce que », inversion sujet-verbe etc).

La troisième colonne aura pour nom *Tag Correct*, comme son nom l'indique ,il sera question de corriger les tags proposés par ALP.

Une approche concernant la détection de questions similaire développée par (Abacha and Dina 2016) repose sur la reconnaissance de *text entailement*,

Cet ensemble de questions équivaut plus ou moins en français à la méthode QQOQCCP: Quoi, Qui, Où, Quand, Comment, Combien, Pourquoi

De cet fait, il nous faudra nous baser sur les recherches sur le *text entailement* d'(Abacha and Dina 2016).

Afin que la machine puisse comprendre ce qui lui est demandé (c'est-à-dire interroger le corpus pour répondre à la question posée par l'utilisateur), il va falloir modéliser la question, c'est-à-dire la faire passer de la langue naturelle à un métalangage compréhensible par le système.

Nous détaillerons en premier lieu les trois paramètres majeurs des systèmes de RAP en nous appuyant sur la classification de ces systèmes effectuée par les chercheurs Saksamudre et al. (Saksamudre, Tech, & Shrishrimal, 2015), à savoir la prise en compte du nombre de locuteurs, la taille du vocabulaire du système et le débit de parole.

Ces systèmes restent les plus simples à développer selon les chercheurs mais restent évidemment les moins flexibles en terme de diversité des locuteurs.

Les modèles langagiers établissent une probabilité d'occurrence d'un mot après une suite de mots, en se basant sur des n-gram, et cela permettra d'associer le son entendu au mot le plus probable dans cette suite de phrase.

Jean-Pierre Jaffré dans *Pourquoi distinguer les homophones ?* (2006), va associer l'homophonie et l'hétérographie, car ces deux phénomènes vont généralement de paire en français.

Dans notre recherche, nous allons analyser des articles scientifiques qui traitent du domaine biomédical et plus précisément des Maladies Tropicales Négligées. Une fois cela fait, nous allons créer une carte pour géolocaliser les MTNs.

L'équipe de chercheurs composée de [Lossio-Ventura et al, 2014] ont combiné une méthode statistique et une méthode de fouille de données afin d'extraire et de classer des termes biomédicaux.

En effet dans leur travail, ils mentionnent l'utilisation de TF-IDF ou "term frequency - inverse document frequency " qui mesure l'importance d'un terme dans un document d'un corpus donné[Lossio-Ventura et al, 2014].

Cet alphabet est proche de l'alphabet latin utilisé pour le français, il contient 29 lettres au total dont 8 voyelles et 21 consonnes présentées dans le tableau suivant, avec leurs transcriptions phonétiques.

Il existe des outils performants qui sont adaptés pour chaque domaine d'étude donné. Le GeoMiner est une extension et une évolution de DBMiner*, un outil puissant qui utilise de multiples technologies pour traiter et géolocaliser des données issues de grands corpus.

Dans le projet Global Neglected Tropical Disease Database (GNTD), les chercheurs ont utilisé plusieurs ressources dans la collecte de données

« Geoparsing » est le processus avec lequel nous pouvons identifier les références géographiques dans un texte » [Kemp,2007]. Dans le Traitement Automatique des Langues, c'est ce qu'on appelle Reconnaissance d'Entités Nommées (REN) [Abascal-Mena & López-Ornelas, 2010].

Selon [Fayyad et al, 2002] une visualisation de données peut fournir une vue d'ensemble qualitative des ensembles de données volumineux et complexes, en outre, résumer ces données complexes et volumineuses.

Si par exemple on veut contrôler les homophones de « respecter » dans la phrase « Merci de la respecter », nous obtiendrions (ici « respecter » n'est pas validé comme ambigu car aucun de ses homophones ne produit une autre sens dans ce contexte):
•Merci de la respectez
•Merci de la respectés
•Merci de la respectées
•Merci de la respecté

Nous avons détecté beaucoup d'erreurs d'ordre grammaticale, comme l'erreur trouvée dans l'article 1 où la présence d'une préposition collée au verbe a été détectée par ALP comme étant une EN PERS.

Les SQR font cependant un véritable bond en avant grâce aux conférences TREC (*Text REtrival Conference*)
ce qui n'est évidemment pas correct et ce qui pourrait induire l'utilisateur en erreur.

(Jurafsky and Martin 2017) définissent trois étapes indispensables à la normalisation de texte. Une fois le texte normalisé, il passe par un processus d'analyse qui comprend les étapes suivantes, définies par (Embarek, 2008).

En RAP nous pouvons citer des systèmes tels que les systèmes de dictée vocale, les assistants vocaux personnalisés, ainsi que l'indexation de documents audio que nous expliquerons par la suite à l'aide des travaux de Joseph Mariani (Mariani, 2002).

Selon Saksamudre et al. nous pouvons faire quatre classes de vocabulaire en fonction de leur taille :
On parlera alors de modèles dépendants de l'utilisateur, ou « speaker dependent models », de modèles indépendants de l'utilisateur, ou « speaker indenpendent models », et de modèles s'adaptant à l'utilisateur, « speaker adaptive models ».

En général, selon Mariani (Mariani, 2002) le découpage se fait à partir d'un corpus vocal.
Selon le chercheur, la plupart des systèmes de RAP utilisent les MMC pour effectuer cette tâche de mise en parallèle.

Nous avons analysé cet exemple extrait de notre corpus d'expérimentation, à l'aide de l'outil de reconnaissance d'entités nommées de Stanford CoreNLP. Le résultat est illustré sur la « Figure 1 » ci-dessous.

Une revue en Open Access existe, PLOS NTD, dédiée spécifiquement à ce sujet.

Les articles du corpus étudié sont écrits en XML : un langage facile à lire par l'homme et la machine.
De ce fait, selon le contenu des phrases, une catégorie est attribuée pour chaque localisation, ainsi nous trouvons par exemple : "The same trend has also been reported in several<LOCATION>European</LOCATION> cities following the fox population growth after the successful vaccination campaign against rabies [76], [145]." [DELHOTAL, 2016]

Cette approche expliquée par Saksamudre et al., est une des premières existant dans le domaine de la RAP et a permis de grandes avancées dans celui-ci.

ALP ne reconnait pas tous les noms de personnes non arabe, nous avons détecté quelques erreurs si le nom de personne est Berbère ou étranger.

Ecriture de paragraphes

Intégrez les deux articles supports (*WordNet* ou *FrameNet*) à votre espace Zotero, puis rédigez un à deux paragraphes qui les synthétise, en utilisant les fonctions de Zotero pour ajouter la référence bibliographique et une bibliographie.

Ecriture de l'état de l'art

Recherchez au moins trois articles sur une thématique au choix (ci-dessous quelques idées de thématiques), puis rédigez un état de l'art faisant la synthèse de ces articles. Vous utiliserez Zotero pour intégrer les références bibliographiques à votre texte.

Thématiques :

- Word Embeddings
- Synthèse ou reconnaissance vocale (*Speech synthesis, speech recognition)*
- Recherche d'information (*Information extraction, data mining*)
- Topic Modelling
- Langues inventées
- Semantic Role Labelling
- Morphologie
- Exploration Contextuelle
- Théorie Sens-Texte
- …

Ecriture de la conclusion et de l'introduction

A partir des articles que vous avez sélectionnés, rédigez une conclusion, puis une introduction.