Détection de Sauts stylistiques et de sauts qualitatifs dans les corpus - Quels rapprochements entre les deux tâches ?

Sorbonne Université des lettres - Master SDL, mention L&I, première année

27/06/2023

Description de la tâche

Plan

Description de la tâche

Format de données et accès à l'information

Comment détecter les variations au sein d'un texte ou d'une collection de textes

Corpus d'études : deux matériaux de nature différente

Méthodes d'analyse de la variation

Résultats

Conclusion

Altération des données

Perte d'informations textuelles

- Erreurs de retranscription
- Retranscription automatique (OCR)
- Problème d'encodage

Perte d'informations extra-textuelles accompagnent le texte

- ToC
- Police
- Fonte
- ...

Limite du support Tous les supports ne peuvent pas représenter les mêmes données

- Structure complexe
- Hyperliens
- Contenu intégré

Une structure textuelle difficilement extractible

Documents nativement numériques

- Format numérique en PDF
- Format textuel en apparence facile d'accès
- Contraintes liées aux documents numériques

Découpage en paragraphes et extraction de structure textuelle

- Découpage en paragraphes moins précis
- Trouver une structure textuelle pour mieux extraire les sections

Description de la tâche

Format de données et accès à l'information

Comment détecter les variations au sein d'un texte ou d'une collection de textes

Corpus d'études : deux matériaux de nature différente

Méthodes d'analyse de la variation

Résultats

Conclusion

Comment mesurer le bruit ?

- Méthodes supervisées :
  - Fiabilité +++
  - Données pré-annotées (inter-annnotation)
  - CER / WER
- Méthodes non-supervisées : Évaluer sans avoir la réponse
  - Fiabilité -
  - Approches statistiques (n-grammes) et utilisation de ressources externes (lexiques)
- Pourquoi utiliserons-nous des méthodes non-supervisées ?

Comment mesurer le bruit ?

- Méthodes supervisées :
  - Fiabilité +++
  - Données pré-annotées (inter-annnotation)
  - CER / WER
- Méthodes non-supervisées : Évaluer sans avoir la réponse
  - Fiabilité -
  - Approches statistiques (n-grammes) et utilisation de ressources externes (lexiques)
- Pourquoi utiliserons-nous des méthodes non-supervisées ?
  - Donnée absente : $\emptyset$ vérité de terrain (jeux de données réels souvent sans car coût de la production)

Comment mesurer le bruit ?

- Méthodes supervisées :
  - Fiabilité +++
  - Données pré-annotées (inter-annnotation)
  - CER / WER
- Méthodes non-supervisées : Évaluer sans avoir la réponse
  - Fiabilité -
  - Approches statistiques (n-grammes) et utilisation de ressources externes (lexiques)
- Pourquoi utiliserons-nous des méthodes non-supervisées ?
  - Donnée absente : $\emptyset$ vérité de terrain (jeux de données réels souvent sans car coût de la production)

Comment caractériser les erreurs sans la correction ? Comment retrouver les parties sans structure?

Détecter des variations de style

- Syntaxe : Les n-grammes d'étiquettes morphosyntaxiques pour l'analyse de récurrences de motifs
- Taux de lexicalité de chaque section
- Observer le rapprochement ou l'éloignement des sections avec une méthode non-supervisée : le *clustering*

Description de la tâche

Format de données et accès à l'information

Comment détecter les variations au sein d'un texte ou d'une collection de textes

Corpus d'études : deux matériaux de nature différente

Méthodes d'analyse de la variation

Résultats

Conclusion

Explosion de la presse en France au 17^{e} siècle, les Mazarinades

- *Corpus or not corpus ?*
  Corpus Antonomaz [1].
  - Textes issus de la Fronde (1648-1653)
  - Baisse des coûts de la presse -> Production massive de courts écrits
    - récits
    - actualités
    - blâmes
    - apologies
    - ...
  - Textes très viraux et diversifiés.

Explosion de la presse en France au 17^{e} siècle, les Mazarinades

- *Corpus or not corpus ?*
  Corpus Antonomaz [1].
  - Textes issus de la Fronde (1648-1653)
  - Baisse des coûts de la presse -> Production massive de courts écrits
    - récits
    - actualités
    - blâmes
    - apologies
    - ...
  - Textes très viraux et diversifiés.
- Qualité du support d'origine très variable
  - Souvent imprimé sur papier fin
  - Conditions de conservation
  - Pages parfois très denses, économie de place?
  - Styles d'écriture très variés

Mazarinades II : Le retour du roi

- Corpus hétérogène
  - En nombre de pages : de une à +500 pages (majorité 4)

Mazarinades II : Le retour du roi

- Corpus hétérogène
  - En nombre de pages : de une à +500 pages (majorité 4)
  - En source : Bibliothèque mazarine, Gallica, Google Livres

Mazarinades II : Le retour du roi

- Corpus hétérogène
  - En nombre de pages : de une à +500 pages (majorité 4)
  - En source : Bibliothèque mazarine, Gallica, Google Livres
  - En langue : Majorité = fra, mais lat et all

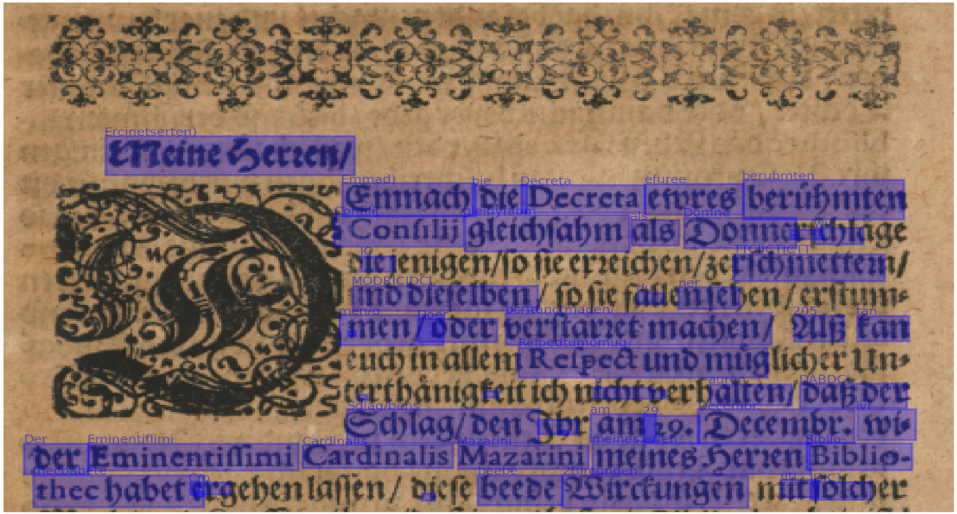
Mazarinades II : Le retour du roi

- Corpus hétérogène
  - En nombre de pages : de une à +500 pages (majorité 4)
  - En source : Bibliothèque mazarine, Gallica, Google Livres
  - En langue : Majorité = fra, mais lat et all
  - En représentation : OCRisés (Kraken) -> rendus disponibles en XML + ODD : fuzzy standard

Mazarinades II : Le retour du roi

- Corpus hétérogène
  - En nombre de pages : de une à +500 pages (majorité 4)
  - En source : Bibliothèque mazarine, Gallica, Google Livres
  - En langue : Majorité = fra, mais lat et all
  - En représentation : OCRisés (Kraken) -> rendus disponibles en XML + ODD : fuzzy standard
- Hétérogénéité perceptible –> chaîne de traitements affectée, mesures compliquées à généraliser
  - Taille variée: TTR, hapax ratio
  - Langue: n-grammes, lexicalité, . . .
  - Sources de qualité très variable
  - "standard" XML/ODD et *parsing*

Accessibilité à l'information



Conférences TALN et RECITAL

- Traitement Automatique des Langues Naturelles (TALN) créée par Philippe Blache en 1994
- Une conférence annuelle depuis 1997
- TALN et Rencontre des Etudiants Chercheurs en Informatique (RECITAL) organisés conjointement depuis 2002
- Conférences internationales TALN destinées aux chercheurs et professionnels du TAL
- Conférences RECITAL principalement destinées aux travaux d'étudiants souvent doctorants

Caractéristiques des corpus TALN et RECITAL

- Disponibles sur un Gitlab d'un membre du comité permanent de TALN
- Des articles en format PDF
- Mises à jour de la ressource courantes

Pour la version 1997 - 2022 :

Conférence	Caractères	Phrases	Mots	Documents
RECITAL	10 047 222	60 740	1 526 182	318
TALN	42 605 195	262 920	6 591 748	1545

Table: Statistiques des articles des deux conférences TALN et RECITAL rédigés de 1997 à 2022 avec le format TXT

Retranscription automatique - OCR

Reconnaissance Optique de Caractères (OCR) :

- Document physique / image –> Document numérique
- Contenu altéré, on passe d'une suite de glyphes positionnés à un endroit précis de la page à une suite de caractères, parfois dans une certaine police.
- Étapes de transformation, faisant des choix (l ou f, espace ou non)
  –> Chaque choix = erreur potentielle

Comment savoir ce qui devrait être écrit?

Retranscription automatique - OCR

Reconnaissance Optique de Caractères (OCR) :

- Document physique / image –> Document numérique
- Contenu altéré, on passe d'une suite de glyphes positionnés à un endroit précis de la page à une suite de caractères, parfois dans une certaine police.
- Étapes de transformation, faisant des choix (l ou f, espace ou non)
  –> Chaque choix = erreur potentielle

Comment savoir ce qui devrait être écrit?

On triche : utilisation de mesures non-supervisées, indépendantes de la vérité de terrain.

Mesures non-supervisées de la qualité d'une retranscription

Aperçu des méthodes explorées

TTR : # de mots différents / longueur du texte (en mots)

Hapax ratio : # d'hapaxes / longueur du texte (en mots)

Hypothèse bruit = plus grand vocabulaire ou plus d'hapaxes

Problème taille des textes, langues

Taux de lexicalité : # d'occurrences dans le lexique / longueur du texte, problèmes :

- Agrégation, pas de détection de bruit local
- Lexique qui couvre les langues + états de langue

Clusterisation à partir du Taux de lexicalité (T_{lex})

- Par page, réduction à 2 dimensions
- Trouver le bon algorithme et # de *clusters* -> WCSS, Silhouette
- Identifier le contenu –> Analyse manuelle, notation, attribution de scores aux *clusters*

Conversion des articles

Conversion des articles en format TXT:

- Utilisation de l'outil pdftotext pour la conversion des PDF en TXT
- Découpage des articles en paragraphes
- Format optimal pour le découpage en paragraphes mais pas pour extraire du contenu structuré
- Des erreurs de conversions dans ce format
- Contenu de chaque paragraphe limité par un flot de caractères

Quelle solution ?

- Générer les articles en XML avec GROBID

Conversion des articles en XML

Conversion des articles en format XML:

- Découpage des articles en sections
- Format adapté pour du contenu structuré
- Des erreurs d'extraction mais peu d'erreurs comparée à la méthode précédente
- Extraction de blocs plus facile avec *BeautifulSoup*

Avantage : Un format qui respecte la structure du document

Détecter la variabilité stylistique des sections

Taux de lexicalité

- Regarder le taux de lexicalité de chaque section

Métriques de calcul de distances

- Scores de similarité (cosinus) : sections successives / toutes les sections

Regroupement des sections par *clustering*

- *Clustering* avec *K-Means* par mots
- Choisir un nombre de *clusters* adapté aux échantillons

Observation de la récurrence des n-grammes d'étiquettes morphosyntaxiques

- Trouver le motif le plus fréquent de chaque section
- Identifier le motif plus fréquent présent dans deux sections

Clustering à partir du T_{lex}

Vecteurs T_{lex} bruts, écarts de T_{lex}, combinaison

Algorithmes Plusieurs testés, *k-means* privilégié

 # *clusters* WCSS. Silhouette score (meilleur) -> 2 infos à rechercher :

- Homogénéité taille *clusters*
- Représentativité de chaque point

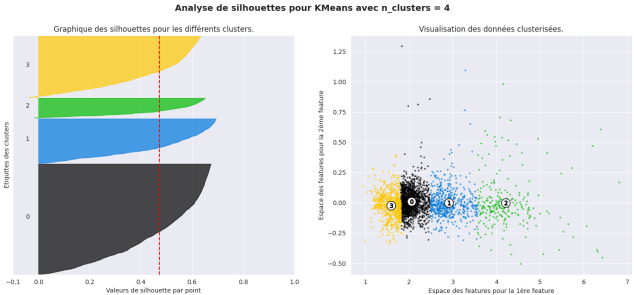
Clustering à partir du T_{lex}

Vecteurs T_{lex} bruts, écarts de T_{lex}, combinaison

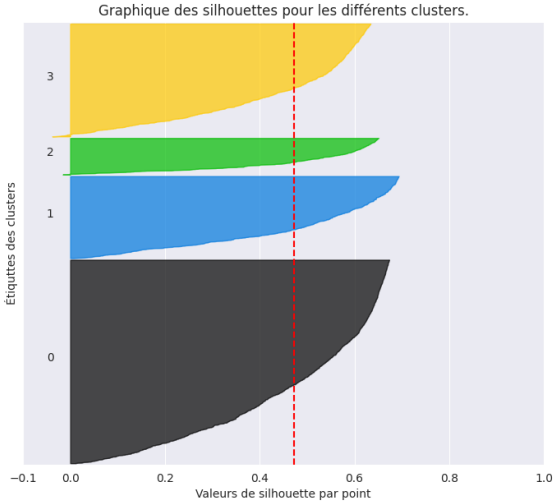
Algorithmes Plusieurs testés, *k-means* privilégié

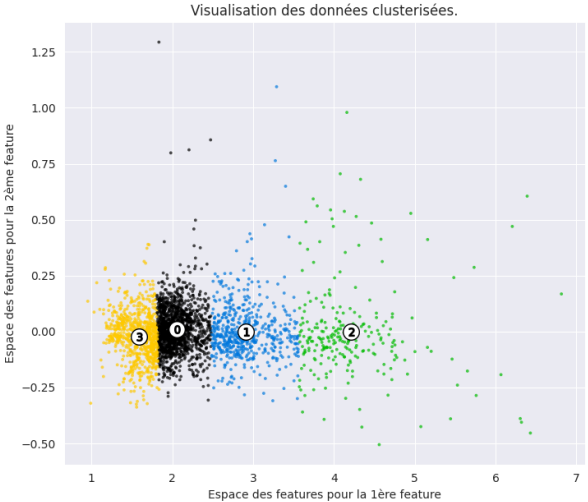
 # *clusters* WCSS. Silhouette score (meilleur) -> 2 infos à rechercher :

- Homogénéité taille *clusters*
- Représentativité de chaque point



Analyse de silhouettes pour KMeans avec n_clusters = 4





Disctimination des *clusters*

6 *clusterings* gardés : Analyse manuelle de quelques textes / *cluster*.
Pas de retranscription complète, mais note /10, quelques textes par *cluster* généré.

Agrégation des notes par *cluster* –> Note du *cluster*.
Pour chaque texte, note = moyenne des notes des *clusters* auxquel il appartient.

257 textes avec note ≤ 4 –> Textes particulièrement bruités.
Évaluation plus globale compliquée.

Résultats des méthodes explorées pour la détection de variations stylistiques entre sections

- Avec un seuil du T_{lex} supérieur ou inférieur à 0.5, on détecte facilement les changements des variations entre sections
- Un *clustering* qui isole plus souvent la première et dernière section.
- Le taux de lexicalité et le score de similarité cosinus concordent très souvent sur des courbes.
- On trouve des intersections d'étiquettes morphosyntaxiques reprises dans des paires de sections de n-grammes où $n = 5$.

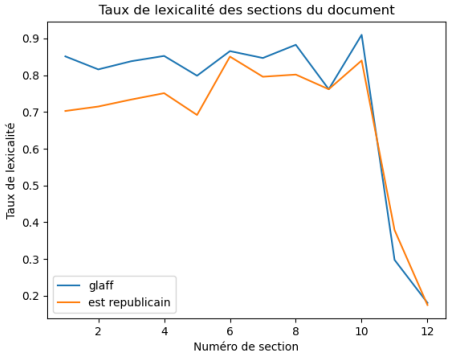


Figure: Variations du taux de lexicalité des sections - article TALN

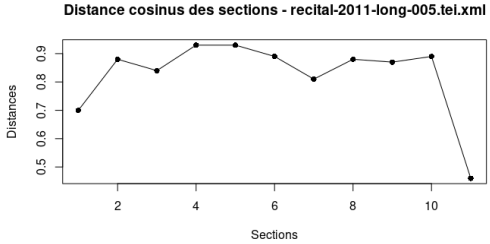


Figure: Scores de similarité cosinus pour les sections - article RECITAL

Description de la tâche

Format de données et accès à l'information

Comment détecter les variations au sein d'un texte ou d'une collection de textes

Corpus d'études : deux matériaux de nature différente

Méthodes d'analyse de la variation

Résultats

Conclusion

Taux de lexicalité

++ Mesure efficace
– Nécessite lexique adapté, dépendant de la langue

Clustering : k-means

++ Rapide, donne de bons résultats
– Trouver le bon nombre de *clusters*

Évaluation - Méthode non-supervisée

–> ∅ vérité de terrain = éval compliquée à généraliser *corpus-wide*
–> Obligés d'analyser les résultats manuellement

Ouverture

Compression Compression du texte dépendant des motifs trouvables [2], hypothèse : meilleur taux de compression = plus de motifs trouvés

Zones de texte Utiliser la détection du bruit, notamment selon la ligne, pour détecter les erreurs de zones de texte

Plongement lexical Un espace sémantique entre les mots pour mieux rapprocher ou éloigner certaines sections

Segments répétés Avec une liste de segments répétés définie dans le projet Hatier 2016 [3], possibilité d'étudier la distribution de ces segments entre sections

Bibliographie

[1] K. Abiven, G. Lejeune, A. Bartz, *et al.*, *Antonomaz/corpus: Collection de mazarinades encodées en xml-tei.* https://github.com/Antonomaz/Corpus. [Online]. Available: https://github.com/Antonomaz/Corpus.

[2] C. Martineau, "Compression de textes en langue naturelle," Theses, Université de Marne-la-Vallée, Dec. 2001. [Online]. Available: https://hal.science/tel-02076650.

[3] S. Hatier, "Identification et analyse linguistique du lexique scientifique transdisciplinaire. Approche outillée sur un corpus d'articles de recherche en SHS," Theses, Université Grenoble Alpes, Dec. 2016. [Online]. Available: https://theses.hal.science/tel-01690554.