Indexer, comparer, apparier des textes et leurs résumés : une exploration.

.

Martine Cadot[(1, 2),] *Sylvain Aubin* [(3)], *Alain Lelu* [(2, 4, 5)]

(1) – Université Nancy1 (UHP)

(2) - LORIA, Nancy

(3) Diatopie S.A., Paris

(4) ISCC, Paris

(5) - Université de Franche-Comté

alain.lelu@univ-fcomte.fr

Aperçu

- Le problème choisi : apparier résumés et textes tronqués.

- 1) Aboutir à une matrice de dissimilarités (textes × résumés)
  - Sans indexation : distances/dissimilarités de compression
  - Avec indexation :
    - par chaînes de caractères
    - par lemmes et expressions composées

  et distances :
    - euclidiennes entre vecteurs-textes normalisés
    - TF-IDF
    - de Hellinger

- 2) Apparier = traiter une matrice de dissimilarités (textes × résumés).

Une méthode en 4 phases.

- Résultats

- Conclusions, perspectives

Défi DEFT 2011 : le problème choisi

- Apparier 300 résumés avec 300 articles de sciences humaines tronqués de leur introduction et conclusion (piste 2.1).
- Difficultés : tâche pas évidente pour certains résumés, même pour des humains spécialistes (parfois : même revue, même auteur, même sujet...).
- Facilités :
  - correspondance biunivoque résumés-textes,
  - autre piste 2.2 plus facile : appariement résumés-textes complets, ET mêmes résumés et textes, non utilisée.
- Intérêt : banc d'essai pour tester
  - différentes distances,
  - différentes stratégies d'appariement,
  - et envisager l'extension à un univers « ouvert » plus réaliste (avec résumés sans textes, textes sans résumés).

Méthode sans indexation : distances/dissimilarités de compression

– Distance de compression (Cilibrasi 2003) entre 2 textes x et y :

$D(x,y) = (Zip(xy)-min(Zip(x),Zip(y))) / max(Zip(x),Zip(y))$

où xy est la concaténation de x et y

Zip(x) est la taille du fichier x après compression

– En pratique, pour des résumés r courts, et des textes t longs :

$D(r,t) = (Zip(rt)-Zip(r)) / Zip(t)$

– Nota 1 : pour la compression par déflation (Zip.exe)

$Zip(tr) \neq Zip(rt)$ → 2 variantes

– Nota 2 : on a utilisé aussi la variante (= dissimilarité) plus « sensible »

$D(r,t) = (Zip(rt)-Zip(t)) / Zip(r)$

– Nota 3 : la combinaison des résultats des 2 × 2 variantes donne les meilleures performances sans indexation (.93)

Méthode avec indexation (1) : basique, par chaînes de caractères

- ~26 000 formes différentes, hors hapax, pour l'ensemble d'apprentissage

- ~21 000 formes différentes, hors hapax, pour l'ensemble de test

Méthode avec indexation (2) : élaborée, par lemmes et expressions composées

- NeuroNav = logiciel de navigation dans une base de textes, selon les axes : mots, documents, thèmes

- On a utilisé ici deux modules de pré-traitement :
  - Étiqueteur simple (verbe, adjectif, nom), avec élimination des mots grammaticaux
  - Extracteur d'expressions composées (patrons syntaxiques + dictionnaire d'exceptions)

→ ~26 000 termes différents (dont ~7000 composés), hors hapax, pour l'ensemble d'apprentissage

~16 000 termes différents, hors hapax, pour l'ensemble de test

Méthodes avec indexation : distances

- **Distances euclidiennes entre vecteurs-textes normalisés**
  - On normalise les vecteurs ($\in$ hypersphère unité)
  - On calcule la distance de la corde : $\sqrt{2}\ (1 - <\mathbf{v}_t, \mathbf{v}_{t'}>)^{1/2}$

- **1) Distance euclidienne (basique) entre vecteurs-textes normalisés**

$\mathbf{v}_t = \{ k_{it} \}$ où k_{it} est l'occurrence du terme i dans le texte t

$\mathbf{v}_t \rightarrow \underline{\mathbf{v}}_t = \{ k_{it} / || \mathbf{v}_t || \}$ où $|| \mathbf{v}_t || = (\Sigma_i\ k_{it}^2)^{1/2}$

Nota : $\mathbf{v}_t$ et $\underline{\mathbf{v}}_t$ sont colinéaires

- **2) Distance TF-IDF**

$\{ k_{it} \} \rightarrow \{ k_{it} \log(N / n_i) \}$ + normalisation

où n_i est la somme des *présences* du terme i dans les textes

Nota : mal adaptée à l'arrivée incrémentale de nouveaux vecteurs.

Méthodes avec indexation : distances (2)

- **3) Distance de Hellinger**
  normalisation : $\{ k_{it} \} \rightarrow \{ (k_{it} / k_{.t})^{1/2} \}$
  D_H = distance de la corde

Propriétés :

-- Liée à l'entropie de Renyi d'ordre ½ entre 2 distributions :

$I^{(1/2)}(\mathbf{x}_q / \mathbf{x}_p) = - 2 \log_2 (\cos(\mathbf{x}_p, \mathbf{x}_q)) = - 2 \log_2 (1 - D_H^2/2)$

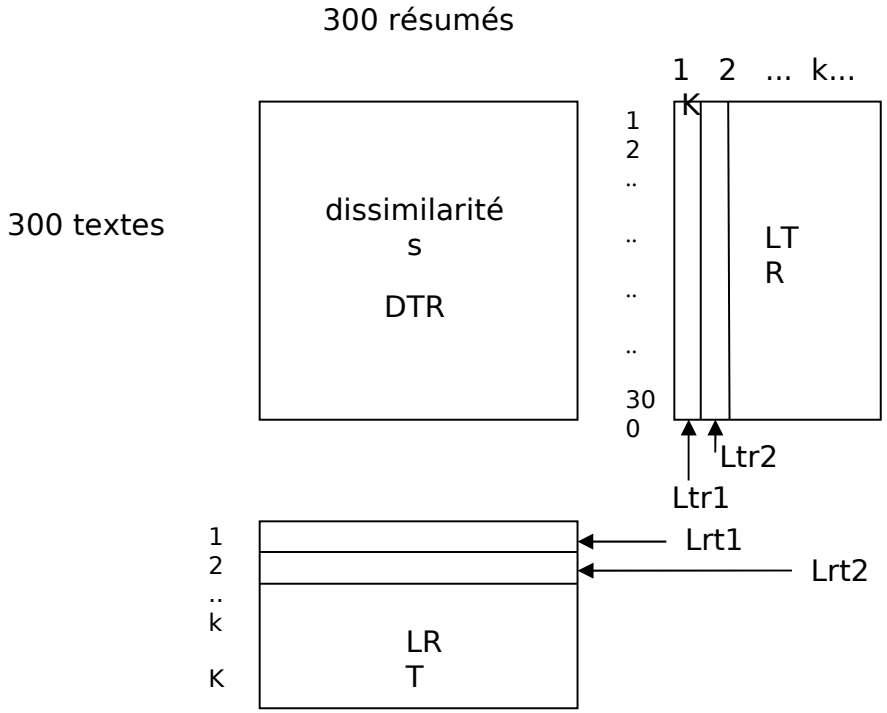
- Equivalence distributionnelle – comme la distance du khi2 (AFC).

- Adaptée à l'arrivée incrémentale de nouveaux vecteurs.

Méthode d'appariement (1)

Illustrée sur l'exemple de la distance de compression (dissim. 2, résumé-puis-texte)

- *Notations* : DTR = matrice des dissimilarités
  - LTR et LRT = N°ID des K plus proches voisins. Ici K=10
  - Ltr1, ou Ltr2, ... définit implicitement une liste de couples (texte, résumé) (idem pour Lrt1, Lrt2, ...)



Méthode d'appariement (2)

- *Etape 1 – Voisins réciproques*
  - Ltr1 et Lrt1 → liste $L^{(+++)}$ de couples communs, de « qualité$^{+++}$ ». Ici un seul couple commun.
- *Etape 2 – Dédoublonnage :* certains résumés (resp. textes) sont plus proches voisins de *plusieurs* textes (resp. résumés). La colonne Ltr1 (resp ligne Lrt1) contient donc des N° de rang « doublons ». Ici Ltr1 contient 298 doublons (295 textes pour un résumé et 3 pour l'autre !) et Lrt1, 48 (22 résumés, chacun pour 2 ou 3 textes). On constitue 4 listes à partir des couples concernés :

- $Ltr1^{(++)}$ est constituée en retirant à Ltr1 :
  - Les couples de $L^{(+++)}$. Ici 1 couple
  - Les couples doublons. On nomme alors $ltr1^{(-)}$ la liste des textes avec résumés doublons. Ici 2 résumés dans cette liste, correspondant à 298 couples.
  - Les couples « contradictoires », c'est à dire dont un des éléments se trouve avec un partenaire différent dans Lrt1. Ici aucun couple.

  → Ici 1 couple dans la liste $Ltr1^{(++)}$.
- Et de même pour $Lrt1^{(++)}$ et $Lrt1^{(-)}$. Ici on a 251couples dans $Lrt1^{(++)}$ et 48 dans $lrt1^{(-)}$.
- La concaténation des deux listes $Ltr1^{(++)}$ et $Lrt1^{(++)}$ donne la liste $L^{(++)}$ des couples de « qualité$^{++}$ », tous différents par construction. Ici 252 couples dans la liste $L^{(++)}$.

Méthode d'appariement (3)

- *Etape 3 – Distinguer vrais et faux plus proches voisins*
  - Parmi les textes de $ltr1^{(-)}$ associés à un même résumé, on choisit celui tel que son écart de distance entre ce résumé et le résumé de Ltr2 soit maximal.

≡ *en cas de doublons, le bon résumé aura toutes chances de « se détacher le plus du peloton de ses poursuivants ».*

→ liste $Ltr1^{(+)}$ de cardinal $|ltr1^{(-)}|$. =2

Et de même une liste $Lrt1^{(+)}$ de cardinal $|lrt1^{(-)}|$. =22

  - On fusionne alors ces deux listes de couples :
    - s'il existe des couples communs aux 2 listes (1 seul ici) on nomme $L^{(+)}$ leur liste, de « qualité$^{+}$ », peu élevée.
    - s'il existe des couples sans partenaires dans la liste opposée ET sans contradictions, on nomme $L^{(0)}$ leur liste, de « qualité $^{(0)}$ », médiocre (ici 20 car 2 couples en contradiction) .
  - *Nota : $L^{(++)}$, $L^{(+)}$, et $L^{(0)}$ ne peuvent se contredire par construction.*

Méthode d'appariement (4)

- *Etape 4 – « Ramasse-miette » pour les résumés non encore appariés (ici 26).*

  On part des tableaux complets LTR et LRTdes K plus proches voisins (empiriquement ici K optimal = 10)

  - On associe à chaque texte le résumé de rang k pour lequel le *saut de dissimilarité* entre k et k+1 est le plus grand.
  - Et de même pour chaque résumé

  – On concatène ces 2 listes et on supprime :
    - Les couples en contradiction dans cette liste,
    - Les couples contradictoires avec $L^{(+++)}$, $L^{(++)}$, $L^{(+)}$,
  – On ajoute la liste résultante à $L^{(0)}$. (Devient de taille 24)

  Nota : cette étape s'est avérée sans objet pour les distances TF-IDF et Hellinger, où tous les résumés se sont trouvés appariés à l'étape 3.

Pour résumer : étape 1) critère des voisins réciproques, 2) élimination des doublons, 3) leur résolution par critère de saut maximal de distance aux 2^e voisins, 4) rattrappage du reste par critère de saut maximal de distance aux K^e voisins .

Résultats (1)

- Méthodes avec compression
  - Chaque variante (dist.1×dissim.2 ; res.-puis-texte × texte-puis-res.) a à l'étape 2 :
    - Très peu d'erreurs
    - Une bonne quantité d'indéterminés
  - On les combine par la règle de décision
    - *Si contradiction ou indétermination : couple non validé*
    - *Si 1 seule décision, ou 2 à 4 accords : couple validé*
  - Résultat : 22 erreurs sur 300, soit 93% de reconnaissances correctes.

Résultats (2)

- Méthodes avec indexation
  - Sur l'ensemble d'apprentissage, 3 variantes produisent un maximum de couples de qualité+++:
    - Formes brutes + TF-IDF (265)
    - Tous lemmes et expressions + Hellinger (265)
    - Noms lemmatisés et expressions + Hellinger (269)
  - Au bout du processus, cette dernière produit 300 couples corrects (100% de reconnaissance correcte), les deux autres 296 et 298.
  - Ces 3 variantes sont retenues pour l'ensemble de test : elles produisent une décision respectivement sur 198 (100%), 197 et 196 couples. Leur cohérence entre elles conduit à notre unique réponse, gagnante.

Conclusion, perspectives

- Tâche en partie cognitivement difficile résolue à 100% sur l'ensemble de test par une méthode utilisant une qualité moyenne de TAL

- ... et à 88% sans TAL du tout ! (marge de progression possible avec de meilleures méthodes de compression)

- Confirmation de la supériorité (légère) de la distance de Hellinger sur TF-IDF, tout en étant compatible avec l'arrivée incrémentale de documents.

- Confirmation des choix de NeuroNav en matière de distance sémantique :
  - Lemmes des noms + termes composés
  - Distance de Hellinger

- Méthode d'appariement généralisable en univers ouvert (textes sans résumés, et inversement) ; marge d'accroissement de performances en intégrant une meilleure qualité d'étiquetage morpho-syntactique et d'extraction de termes composés.

Merci pour votre attention !