Linguistique de corpus

Sémantique lexicale

Patrick Paroubek

LIMSI-CNRS
Dépt. CHM - Groupe LIR
Bât. 508 Université Paris XI, 91403 Orsay Cedex
pap@limsi.fr

mercredi 19 avril 2017 / Semestre 2 - Cours 9

Source

Ce cours s'inspire et reprend quelques transparents du cours "Acquisition de connaissances sémantiques lexicales en corpus", de Benoît Habert.

Recherche documentaire
Évaluation

Recherche documentaire : trouver un document en réponse à une question.
Au départ un problème de bibliothécaire, résolu au moyen d'indexes affectés manuellement, devenu depuis avec les documents électroniques un problème de moteur de recherche.
Un humain est une entité ancrée et située dans le monde, deux propriétés qui contribuent à l'apprentissage du langage (Wittgenstein). La machine n'a aucune de ces 2 propriétés, elle ne peut s'appuyer que sur la graphie (forme) des mots, mais pas leur sens.

Évaluation

- **bruit** documents non-pertinents retournés, mesure complémentaire la **précision**
- **silence** document pertinents non-retournés, mesure complémentaire le **rappel**

Les mesures de précision/rappel ont été mises au point en 1960 [C.W. Cleverdon, *The ASLIB Cranfield research project on the comparative efficiency of indexing systems*, ASLIB Proceedings, 1960, vol. 12, pp 421-431, ISSN : 0001-253X / DOI : 10.1108/eb049778.].

Automatique

Pour répondre automatiquement aux problèmes suivants :

- Trouver un document qui aborde un thème donné.
- Trouver un document qui contient la réponse à une question.

Il faut au moins être capable de répondre à la question :
Étant donné 2 documents, dans quelle mesure est-ce qu'ils parlent de la même chose ?

Types de tâches

2 niveaux privilégiés pour accéder à la représentation du sens (unités linguistiques servant de clés d'indexation) :

1. mots (simples ou composés)
2. énoncé longs, phrase, phrasette, paragraphe, document.

Trois types de tâches :

1. découper (identifier en contexte les frontières les unités linguistiques servante de clé d'indexation)
2. partitionner (définir des classes/catégories sémantiques)
3. répartir les instances des unités linguistiques rencontrés dans les documents sur les différentes classes définies

Exemple
Corpus de comptes rendus de maternité

A = **bébé calme . ne bouge même pas pendant les soins .** semble détendu **.**
B = **bébé calme .** s'agite peu **même pendant les soins . ne bouge pas** beaucoup **.**

A & B =	10	**bébé (1) bouge (2) calme (3) les (4) même (5) ne (6) . (7) pas (8) pendant (9) soins (10)**
A seulement =	2	semble (1) détendu (2)
B seulement =	4	s' (1) agite (2) peu (3) beaucoup (4)
ni A ni B =	3	endormi (1) au (2) repos (3)

Mesurer la similarité
Jaccard

Les traits (présence/absence d'une forme/graphie), en anglais : *feature*.

	présents dans A	absents de A
présents dans B	A &f B (10)	non A & B (4)
absents de B	A & non B (2)	non A & non B (3)

Indice de **similarité de Jaccard** :

$$J(A,B) = \frac{taille \quad de \quad l'intersection}{taille \quad de \quad l'union} = \frac{|A \cap B|}{|A \cup B|} = \frac{10}{16} = 0.625 \tag{1}$$

$$\begin{aligned} A = B &\Rightarrow J(A,B) = 1 \\ A \neq B &\Rightarrow J(A,B) = 0 \end{aligned} \tag{2}$$

Mesurer la similarité
distance, coefficient de Dice

La **distance de Jaccard** est la valeur complémentaire au coefficient de Jaccard

$$Distance \quad de \quad Jaccard \quad (A, B) = 1 - J(A, B) = 0.375 \qquad (3)$$

Mesurer la similarité
distance, coefficient de Dice

Le coefficient de Dice ressemble beaucoup au coefficient de Jaccard, il varie comme lui entre 0 et 1.

$$D(A,B) = \frac{2 \quad fois \quad la \quad taille \quad de \quad l'intersection}{la \quad somme \quad des \quad tailles} = \frac{2\times|A\cap B|}{|A|+|B|} \qquad (4)$$
$$D(A,B) = \frac{2\times 10}{12+14} = 0.76$$

Mais il accorde plus de poid à la partie commune,
$J(A,B) = 0.625$ mais $D(A,B) = 0.76$.

Mesurer la similarité
conclusion

- Perte de l'enchaînement des mots, approche **sac de mots** (en anglais *bag of words*).
- il existe de **très nombreuses mesures de similarités**, Jaccard, Dice, etc.
- **paramétrage** de la comparaison :
  1. choix des **traits** (tokens, graphies/formes, occurences, mots, mots + parties du discours, lemmes...)
  2. choix de la **mesure de similarité** (Jaccard, Dice, ...)
  3. **pondération des traits** (par ex. accorder plus de poids aux noms, aux mots peux fréquents, ...)

Pondération des traits
tf.idf

Pour pondérer les traits, en Recherche d'Information on utilise souvent la mesure **tf.idf** frequence du terme par la fréquence inverse documentaire (*term frequency . inverse document frequency*) qui renforce le poids des traits discriminants d'un document par rapport aux traits présents dans tous les documents.

$$tf = \frac{le\ \ nombre\ \ d'occurences\ \ du\ \ terme\ \ dans\ \ le\ \ document}{le\ \ nombre\ \ total\ \ d'occurences\ \ du\ \ document}$$

$$idf = \frac{le\ \ nombre\ \ total\ \ de\ \ documents}{le\ \ nombre\ \ de\ \ documents\ \ contenant\ \ au\ \ moins\ \ une\ \ occurence\ \ du\ \ terme} \quad (5)$$

Pondération des traits
tf.idf

La pondération **tf.idf** permet de

1. **caractériser la spécificité**
2. d'un **trait** (token/graphie/forme/morphème/mot/terme/etc.),
3. par rapport **au document** qui le contient,
4. dans le context du **corpus** duquel ce document est extrait.

Expérimentations
Statistiques de base

- Télécharger les 2 ouvrages de Jules Vernes, « Cinq semaines en ballon »et « Les cinq cents millions de la Begum »du site de l'ABU en version texte :
  - `http://abu.cnam.fr/cgi-bin/go?ballon1`
  - `http://abu.cnam.fr/cgi-bin/go?begum2`
- Pour chaque texte créez un répertoire et rangez-y le texte découpé en tronçons de 5 caractères, vous obtenez ainsi un corpus constiuté de petits documents, regroupés dans 2 sous-corpus, un sous-corpus « ballon »et un sous-corpus « begum ».

Correction :

- Téléchargement des version texte non-formaté, puis suppression manuelle du cartouche ABU en début de fichier. Notez que les fichiers sont au format ISO-8859-1
- ```
  > ls /tmp$ ls -l *_jv.txt
  -rw-rw-r-- 1 pap pap 497785 Apr 18 23:32 ballon_jv.txt
  -rw-rw-r-- 1 pap pap 321820 Apr 18 23:34 begum_jv.txt
  > mkdir ballon; mkdir begum
  > cd ballon ; split -b 500 ../ballon_jv.txt
  > ls
  xaa  xbq  xdg  xew  xgm  xic ...
  > cd ../begum ; split -b 500 ../begum_jv.txt
  > split -b 500 ../begum_jv.txt
  ```
  ```

Correction : Transcodage des fichiers de ISO-8859-1 (aussi appelé Latin 1) vers UTF-8 :

```
> cd ../ballon
> for f in `ls -1`
do
  recode latin1..utf-8 $f
done
```

Correction : Notez que les deux documents sont de tailles différentes : 644 fichier pour le sous-corpus begum et 996 pour le sous-corpus ballon.

```
> cd begum; ls -1 | wc -l
644
> cd ../ballon; ls -1 | wc -l
966
```

Expérimentations
Statistiques de base

- Calculer les fréquences des formes « le », « ballon », « armes »dans les deux sous-corpus.

	begum	ballon
nombre occ. total	55881	90299
nombre occ. « le »	996	2073
fréq. rel. « le »	0.01782	0.02295
nombre occ. « ballon »	0	113
fréq. rel. « ballon »	0	0.00125
nombre occ « armes »	4	15
fréq. rel. « armes »	0.00007	0.00016

```
> cd ballon
> cat * | tr ' ' '\012' | sort | uniq -c | sort -k 1,1 | egrep " le$"
  2073 le
```

- Calculer les coefficients de jaccard et dice pour les deux sous-corpus.
- Parmi les trois formes dont nous avons calculé la fréquence relative laquelle est la plus spécifique du sous-corpus `ballon` ?

occ. total	occ. communes
146180	3551

```
> cd ballon;  cat * | tr ' ' '\012' | sort > /tmp/occ_ballon
> cd ../begum;  cat * | tr ' ' '\012' | sort > /tmp/occ_begum
> # occ. communes:
> cat  /tmp/occ_ballon /tmp/occ_begum | sort | uniq -c | egrep " 2 " | wc -l
```

Jaccard	Dice
0.02429	0.04858

La forme la plus spécifique du sous corpus `ballon` est la forme « ballon ».