Indexation d'articles scientifiques
Présentation et résultats du défi fouille de textes DEFT 2016

Béatrice Daille* Sabine Barreaux† Florian Boudin* Adrien Bougouin* Damien Cram* Amir Hazem*
*LINA – UMR CNRS 6241, Nantes
† INIST CNRS, Vandœuvre-lès-Nancy

4 juillet 2016

UNIVERSITÉ DE NANTES

lina LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

TermITH

Indexation automatique

Identifier un ensemble de mots-clés pour un document

Mot-clé

- Mot ou expression
- Représente un sujet important d'un document
- Explicite ou implicite

Tache DEFT 2016

Indexation professionnelle

Identification de mots clés d'articles scientifiques proposés par des indexeurs professionnels

- Cohérence : même concept/même mot-clé. Utilisation de thésaurus
- Exhaustivité : couverture des notions présentes dans l'article

Domaines de spécialités

- linguistique
- sciences de l'information
- archéologie
- chimie

Données

Corpus et référentiels dans chaque domaine

- notices : titre + résumé + mots-clés
- thésaurus

Exemple

La cause linguistique *Linguistique*

L'objectif est de fournir une définition de base du concept linguistique de la cause en observant son expression. Dans un premier temps, l'A. se demande si un tel concept existe en langue. Puis il part des formes de son expression principale et directe (les verbes et les conjonctions de cause) pour caractériser linguistiquement ce qui fonde une telle notion.

Mots-clés : français ; interprétation sémantique ; conjonction ; expression linguistique ; concept linguistique ; relation syntaxique ; cause.

Corpus

Corpus	**Documents**			**Mots-clés**	
	Quantité	Mots moy.	Quantité moy.	« À assigner »	Long. moy.
Linguistique	515	160,5	8,6	61 %	1,7
Sciences de l'info.	506	105,0	7,8	68 %	1,8
Archéologie	518	221,1	16,9	37 %	1,3
Chimie	582	105,7	12,2	76 %	2,2

Corpus : traitements linguistiques

Objectif : encourager les participants à utiliser les mêmes corpus analysés

- segmentation en phrases par PunktSentenceTokenizer - librairie Python NLTK
- segmentation en mots par Bonsai du Bonsai PCFG-LA parser 3
- étiquetage syntaxique réalisé par MElt

Thésaurus

Domaine	**Total entrées**	**Composition**	
		Vocabulaire contrôlé	Volume entrées
Linguistique	13 968	ML (sciences du langage)	6 079
		MC (sciences de l'éducation)	2 681
		MS (sociologie)	5 208
Sciences de l'info.	92 472	MX (Sciences exactes, sciences de l'ingénieur et technologies)	92 472
Archéologie	4 905	MA (art et archéologie)	1 849
		MH (préhistoire et protohistoire)	3 056
Chimie	122 359	MX (Sciences exactes, sciences de l'ingénieur et technologies)	92 472
		M3 (Physique)	29 887

Évaluation

Mesures de la piste 5 de SemEval 2010 - Égalité stricte entre les racines des mots-clés fournies par SNOWBALL.

Mesures

$$\mathrm{P}(d) = \frac{\#\text{nb mots-clés extraits corrects}(d)}{\#\text{nb mots-clés extraits}(d)} \tag{1}$$

$$\mathrm{R}(d) = \frac{\#\text{nb mots-clés extraits corrects}(d)}{\#\text{nb mots-clés de référence}(d)} \tag{2}$$

$$\mathrm{F}(d) = 2 \times \frac{\mathrm{P}(d)\mathrm{R}(d)}{\mathrm{P}(d) + \mathrm{R}(d)} \tag{3}$$

Participants

LIMSI	*Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur : Thierry Hamon*
LINA	*Laboratoire d'Informatique de Nantes Atlantique, Université de Nantes : Adrien, Bougouin, Florian Boudin et Béatrice Daille*
LIPN	*Laboratoire d'Informatique de Paris Nord, Université Paris 13 : Haïfa Zargayouna et Davide Buscaldi*
EBSI	*École de Bibliothéconomie et des Sciences de l'Information, Université de Montréal : Dominic Forest, Jean-François Chartier et Olivier Lacombe*
EXenSa	*SAS eXenSa : Morgane Marchand*

Calendrier

Rappel : 1 piste. 3 méthodes au plus par participant

- Diffusion du corpus apprentissage : 2 mars 2016
- 6 semaines d'entraînement
- Tests sur trois jours entre le 11 et 17 avril 2016

Résultats

Résultats globaux avec les meilleurs scores de chaque équipe

Moy(Préc.)	Moy(Rap.)	**Moy(f-score)**
24.92	30.40	**25.03**

Classement des participants

par points

Rang	Équipe candidate	Points
1er	**eXenSa**	**18**
2ième	EBSI	16
3ième	LINA	12
4ième	LIMSI	7
4ième	LIPN	7

Classement des méthodes

Rang	Méthode	Moy(Préc.)	Moy(Rap.)	**Moy(F-mesure)**
1ier	exensa-m1	28.24	34.37	**29.30**
2ième	ebsi-m2	27.44	33.05	**29.13**
3ième	ebsi-m1	27.73	32.24	**28.88**
4ième	ebsi-m3	25.78	30.85	**27.28**
5ième	lina-m3	30.00	24.67	**26.01**
6ième	lina-m1	28.39	23.53	**24.71**
7ième	limsi-m2	25.75	20.23	**21.65**
8ième	limsi-m1	24.31	21.88	**21.42**
9ième	limsi-m3	25.24	19.79	**21.20**
10ième	lipn-m3	13.28	39.66	**19.04**
11ième	lina-m2	22.21	17.79	**18.91**
12ième	lipn-m1	16.67	21.59	**17.12**
13ième	lipn-m2	14.12	24.03	**17.11**

Classement pour la linguistique

#	Candidat	Préc.	Rap.	**F-mesure**	Points
1.	ebsi-m2	30.26	34.16	**31.75**	5
2.	exensa-m1	23.28	32.73	**26.30**	4
3.	lina-m3	23.16	25.85	**24.19**	3
4.	lipn-m2	13.98	30.81	**19.07**	2
5.	limsi-m2	15.67	16.10	**15.63**	1

Classement pour les sciences de l'information

#	Candidat	Préc.	Rap.	**F-mesure**	Points
1.	ebsi-m1	31.03	28.23	**28.98**	5
2.	exensa-m1	21.26	30.32	**23.86**	4
3.	lina-m3	21.93	21.83	**21.45**	3
4.	lipn-m2	11.72	23.54	**15.34**	2
5.	limsi-m2	13.83	12.01	**12.49**	1

Classement pour l'archéologie

#	Candidat	Préc.	Rap.	**F-mesure**	Points
1.	exensa-m1	43.48	52.71	**45.59**	5
2.	limsi-m3	55.26	38.03	**43.26**	4
3.	lina-m3	53.77	33.46	**40.11**	3
4.	ebsi-m2	30.77	43.24	**34.96**	2
5.	lipn-m1	33.93	31.25	**30.75**	1

Classement pour la chimie

#	Candidat	Préc.	Rap.	**F-mesure**	Points
1.	exensa-m1	24.92	21.73	**21.46**	5
2.	ebsi-m2	19.67	25.07	**21.07**	4
3.	lina-m3	21.15	17.54	**18.28**	3
4.	lipn-m3	10.88	30.25	**15.31**	2
5.	limsi-m2	18.19	14.90	**15.29**	1

Conclusion

Tache classique : simuler l'indexation réalisée par des indexeurs professionnels
Tache difficile : f-mesure moyenne 25,3 %
Écarts entre les domaines