Extracteur Terminologique Statistique

Hugo Larochelle

2002

Plan

- Introduction
- Prérequis à l'extraction
- Erreurs et remarques sur les prérequis
- Quelques mots sur la subjectivité de l'extraction
- Architecture de l'extracteur
- Évaluation des métriques
- Résultats finaux
- Conclusion
- Voies futures

Introduction

Extraction terminologique:

- Qu'est-ce qu'un terme ?
  - *un terme est une représenation littéraire d'un concept dans un domaine donné* [a]
- Pourquoi extraire des termes ?
  - recherche d'information
  - traduction
  - extraction d'information
- Doit-on étudier le sens ?
  - pas nécessairement, car il existe des tests statistiques qui permettent d'évaluer la pertinence d'un terme selon d'autres critères, soit la fréquence et la rareté

[a] voir (Jacquemin, 1997)

Prérequis de l'extraction

L'extraction terminologique et l'évaluation de celle-ci nécessite certains prérequis.

Le modèle de cette ligne de commande Unix les montre tous:

cat {corpus} | {étiquetteur} | {lemmatiseur} | {extracteur}

Prérequis de l'extraction (suite)

Corpus de référence

- corpus sur l'alimentation en eau
  - 12492 mots
  - liste de termes extraits ne contient que des expressions (plus d'un mots)
  - Office de la langue française
  - extraction manuelle et corrections selon la sortie des logiciels
- corpus de médecine
  - 3296 mots
  - liste des termes extraits contient des mots et des expressions
  - membres du RALI/LLI
  - extraction manuelle individuelle, convergence des résultats et ajustements

Corpus	Nb termes $f = 1$	Nb termes $f > 2$
Eau	164	61
Médecine	84	103

Prérequis de l'extraction (suite)

Étiqueteur [a]

Permet de "tokenizer" et d'étiqueter grammaticalement un texte à l'aide d'un lexique. Exemple:

Les	Dete-dart-ddef-masc-plur
enfants	NomC-masc-plur
s'	Pron-prfl-prea-genI-nomI-p3
amuse	Verb-IndPre-sing-p3
dans	Prep
le	Dete-dart-ddef-masc-sing
parc	NomC-masc-sing
.	Punc-pcst
{EOF}	

[a] **voir (Foster, 1991)**

Prérequis de l'extraction (suite)

Lemmatiseur

Permet d'obtenir le lemme de chacun des "tokens" du texte.
Exemple:

Les	Dete-dart-ddef-masc-plur/le
enfants	NomC-masc-plur/enfant
s'	Pron-prfl-prea-genI-nomI-p3/me
amuse	Verb-IndPre-sing-p3/amuser
dans	Prep/dans
le	Dete-dart-ddef-masc-sing/le
parc	NomC-masc-sing/parc
.	Punc-pcst
{EOF}	

Exemple d'entrées du lexique:

industries	NomC	industrie
industriel	NomC	industriel
industriels	NomC	industriel

Erreurs et remarques sur les prérequis

Corpus de référence

- extraction manuelle ne comporte que des expressions pour le corpus de l'eau
- termes extraits sont sous leur forme neutre, et j'ai dû trouver la forme apparaissant dans le texte
- beaucoup trop de termes sont de fréquence unitaire

Erreurs et remarques sur les prérequis

Étiqueteur

- étiquetage est parfois erroné. Exemple:

un	Dete-dart-dind-masc-sing
massif	AdjQ-masc-sing
filtrant	AdjQ-masc-sing

- certains symboles sont associés injustement à des noms communs (%, *, —, etc.);
- segmentation du texte est quelque fois mal réalisée. Exemple:

de pompage	AdjQ-masc-sing
fonctionnel	AdjQ-masc-sing

- mots rares souvent mal analysés grammaticalement Exemple:

antigen	Quan-ndg-sgpl-Sord-ind
antigen	NomC-sing
antigen	Adve-XNOT
antigen	AdjQ

Erreurs et remarques sur les prérequis (suite)

- étiquetage ne peut être fait dans deux langues simultannément. Exemple du corpus de l'eau

These	NomP
problems	NomP
pose	Verb-IndPre-sing-p3
a	Verb-IndPre-sing-p3
considerable	Verb-ParPas-masc-sing
challenge	NomP
to	NomP
water	NomP
utilities	NomP
and	NomP
other	NomP
well	NomP
owners	NomP
in	NomP
North	NomP
America	NomP
and	NomP
around	NomP
the	NomP
world	NomP
.	Punc-pcst

Quelques mots sur la subjectivité de l'extraction

Il suffit d'essayer soi-même d'extraire des termes pour réaliser que la subjectivité est de mise.

L'extraction faite par le RALI/LLI exprime bien ce fait.

- le nombre de termes approuvés par personne varie de 99 à 343
- le tableau suivant montre à quel point le nombre de termes faisant consensus diminue avec le nombre de personne du consensus

Nb personnes	Nb de termes
5	55
4	104
3	187
2	269
1	427

Architecture de l'extracteur

L'extracteur est divisé de la façon suivante:

- Lecture du *corpus monde*
- Lecture du corpus à analyser
- Création du SFX et du LCP
- Recherche des séquences et assignation des scores
- Filtration normale

Lecture du *corpus monde* et du corpus à analyser

Qu'est-ce que le *corpus monde* :

- il permettra de mesurer la rareté d'un mot
- le Hansard a servi de *corpus monde*

Exemple:

2968 attitude

Corpus à analyser

- À l'aide de l'utilisation du SFX (suffixe array) et du LCP (longest common prefixe), il est possible d'obtenir rapidement la fréquence et les occurences de toute séquence apparaissant dans un corpus. Voir (Russell, 1998).

Recherche des séquences et assignation des scores

Différentes variables sont requises par les métriques.

- fréquence **f**
- fréquence mondiale **F**
- variables **a**, **b**, **c** et **d**, permettant de mesurer la liaison entre deux lemmes et définies par le tableau de contingence suivant:

	B	¬B
A	a	b
$\neg A$	c	d

Recherche des séquences et assignation des scores (suite)

L'éventail des métriques testées est très grand. En voici quelques unes:

Pour les mots

- **Entropie (E)**

$$
\begin{aligned}
e(w_1^n) &= (e_{left}(w_1^n) + e_{right}(w_1^n))/2 \\
e_{left}(s) &= \sum_{w|ws \in T} h\left(\frac{|ws|}{|s|}\right) \\
e_{right}(s) &= \sum_{w|sw \in T} h\left(\frac{|sw|}{|s|}\right) \\
h(x) &= -x \log_2(x)
\end{aligned}
$$

Faible entropie

$$
\left.\begin{array}{l} \text{par} \\ . \\ \text{autre} \\ \text{ce} \end{array}\right\} \text{ exemple } \left\{\begin{array}{l} , \\ . \\ 350 \\ 400 \end{array}\right. \quad (22) \left\{\begin{array}{l} \text{l'} \\ \text{en} \\ \vdots \\ \text{d'} \end{array}\right\} \text{ eau } \left\{\begin{array}{l} \text{de} \\ \text{potable} \\ \vdots \\ \text{peut} \end{array}\right\} (117
$$

Forte entropie

Recherche des séquences et assignation des scores (suite)

- **Score de comparaison avec le monde (S)**

$$S = -f \log_2 \left(\frac{f+F}{|T|+|M|} \right)$$

Exemple pour S (corpus de médecine):

monocytes:	$f = 2$ et $F = 0$	$\rightarrow S = 47.7892$
presence:	$f = 2$ et $F = 1796$	$\rightarrow S = 28.1648$

Recherche des séquences et assignation des scores (suite)

Pour les expressions

- **Ratio de vraisemblance (L)**

$$\begin{aligned} L &= a\log a + b\log b + d\log d + N\log N \\ &\quad -(a+c)\log(a+c) - (a+b)\log(a+b) \\ &\quad -(c+d)\log(c+d) - (d+b)\log(d+b) \end{aligned}$$

  où **N** est la taille du corpus. Ce ratio est relativement répandu. À vrai dire, c'est le test de vraisemblance appliqué dans un contexte binomial.

- **Entropie (E)**

  idem à l'entropie pour les mots.

Filtration normale

On filtre finalement à l'aide d'un seuil normal.

Exemple avec seuil $t = 2$

Terme	Entropie	Entropie normalisée	Choisi
puits artésien	49.6853	6.9507	x
eau souterraine	29.9798	3.9923	x
quantité d' eau	9.8399	0.9687	
facture finale	3.4624	0.0113	

Évaluation des métriques

On cherche maintenant à observer le travail fait par chacune des métriques décrites plus haut. Pour ce faire, on compare les métriques avec, entre autre, le bruit et le silence, définies comme suit:

Bruit nombre de termes extraits automatiquement qui ne se trouvent pas dans la liste de référence sur le nombre de termes extraits

Silence nombre de termes non extraits automatiquement et se trouvant dans la liste de référence, sur le nombre de termes dans cette liste

Ces quantités sont exprimées en pourcentage.

Expressions (évaluation)

N	10.00	20.00	30.00	40.00	50.00	60.00	70.00	80.00	90.00	100.00
f	100.00	80.00	73.33	65.00	62.00	63.33	57.14	55.00	54.44	53.00
l	100.00	65.00	66.67	60.00	58.00	51.67	48.57	46.25	48.89	48.00
d	50.00	55.00	56.67	65.00	60.00	55.00	51.43	51.25	53.33	53.00
dm	60.00	65.00	63.33	55.00	50.00	46.67	47.14	45.00	47.78	49.00
fag	50.00	60.00	63.33	67.50	56.00	53.33	52.86	51.25	53.33	53.00
mim	70.00	65.00	56.67	50.00	50.00	46.67	45.71	47.50	47.78	49.00
s	70.00	60.00	66.67	65.00	64.00	58.33	57.14	56.25	52.22	51.00
c	80.00	75.00	70.00	60.00	56.00	53.33	47.14	50.00	52.22	52.00
e	100.00	90.00	70.00	70.00	62.00	58.33	55.71	51.25	50.00	52.00
kuc	50.00	55.00	56.67	65.00	60.00	55.00	51.43	51.25	53.33	53.00
och	50.00	55.00	56.67	65.00	60.00	55.00	51.43	51.25	53.33	53.00
chi	50.00	35.00	43.33	40.00	36.00	35.00	40.00	41.25	42.22	45.00
smc	50.00	55.00	56.67	65.00	60.00	55.00	51.43	51.25	53.33	53.00
phi	60.00	50.00	40.00	40.00	38.00	40.00	41.43	45.00	46.67	48.00
mi	50.00	50.00	46.67	40.00	36.00	38.33	40.00	42.50	46.67	48.00
y	70.00	70.00	53.33	52.50	54.00	53.33	51.43	55.00	52.22	53.00
fa	80.00	80.00	76.67	62.50	62.00	56.67	57.14	52.50	52.22	49.00

Table 1: Progression de la précision des métriques sur le corpus de l'eau pour les expressions

Expressions (suite)

N	3.00	6.00	9.00	12.00	15.00	18.00	21.00	24.00	27.00	30.00
f	100.00	100.00	88.89	91.67	93.33	88.89	85.71	83.33	85.19	83.33
l	100.00	100.00	88.89	91.67	86.67	88.89	85.71	75.00	74.07	73.33
d	33.33	33.33	55.56	66.67	73.33	72.22	71.43	62.50	62.96	60.00
dm	100.00	100.00	88.89	83.33	86.67	88.89	85.71	79.17	81.48	76.67
fag	0.00	50.00	66.67	75.00	80.00	77.78	71.43	66.67	66.67	63.33
mim	66.67	83.33	88.89	83.33	80.00	83.33	71.43	70.83	70.37	66.67
s	100.00	83.33	77.78	75.00	80.00	77.78	76.19	70.83	62.96	66.67
c	100.00	66.67	66.67	50.00	53.33	44.44	42.86	50.00	55.56	60.00
e	100.00	100.00	100.00	91.67	93.33	94.44	90.48	91.67	88.89	83.33
kuc	33.33	33.33	55.56	66.67	73.33	72.22	71.43	62.50	62.96	60.00
och	33.33	33.33	55.56	66.67	73.33	72.22	71.43	62.50	62.96	60.00
chi	66.67	50.00	44.44	41.67	33.33	33.33	33.33	29.17	33.33	40.00
smc	33.33	33.33	55.56	66.67	73.33	72.22	71.43	62.50	62.96	60.00
phi	66.67	66.67	44.44	50.00	60.00	55.56	47.62	41.67	37.04	43.33
mi	66.67	66.67	44.44	50.00	53.33	44.44	38.10	41.67	44.44	46.67
y	100.00	66.67	55.56	50.00	46.67	44.44	52.38	58.33	62.96	63.33
fa	100.00	83.33	77.78	75.00	80.00	77.78	76.19	75.00	66.67	66.67

Table 2: Progression de la précision des métriques sur le corpus de médecine pour les expressions

Expressions (suite)

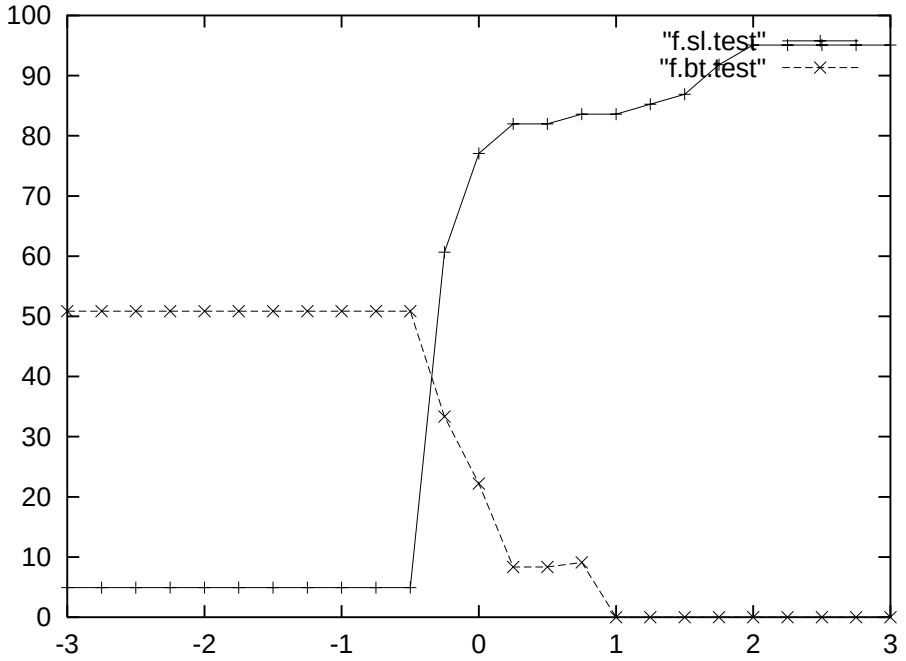


Figure 1: Évolution du bruit et du silence avec la fréquence pour les expressions

Expressions (suite)

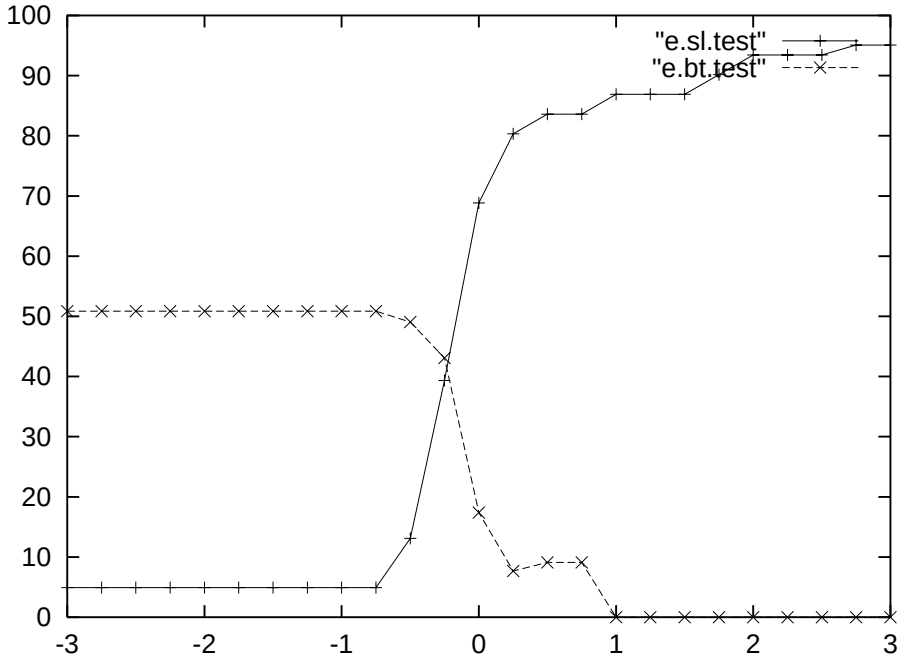


Figure 2: Évolution du bruit et du silence avec l'entropie pour les expressions

Expressions (suite)

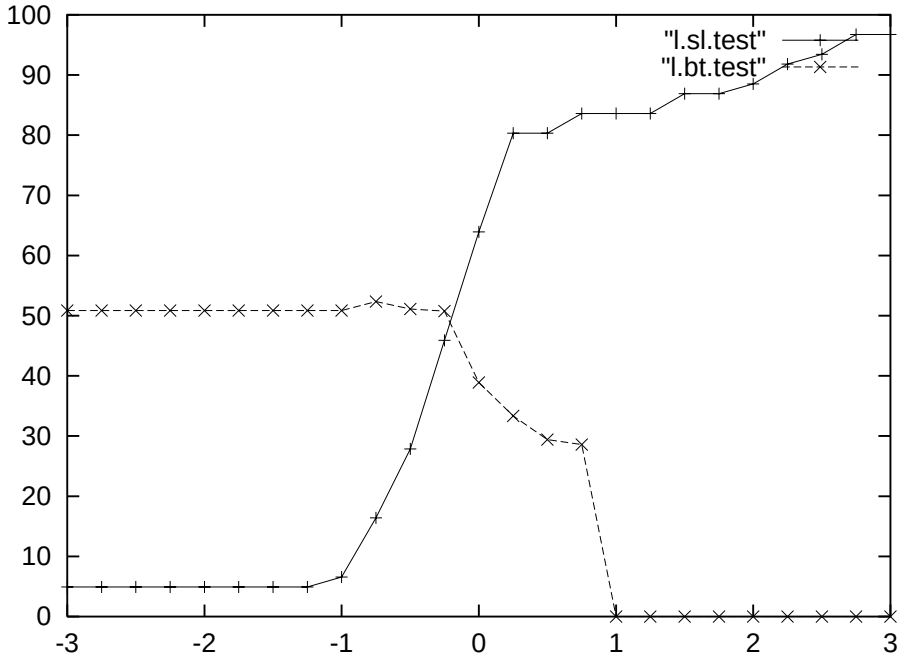


Figure 3: Évolution du bruit et du silence avec le ratio de vraisemblance pour les expressions

Expressions (suite)

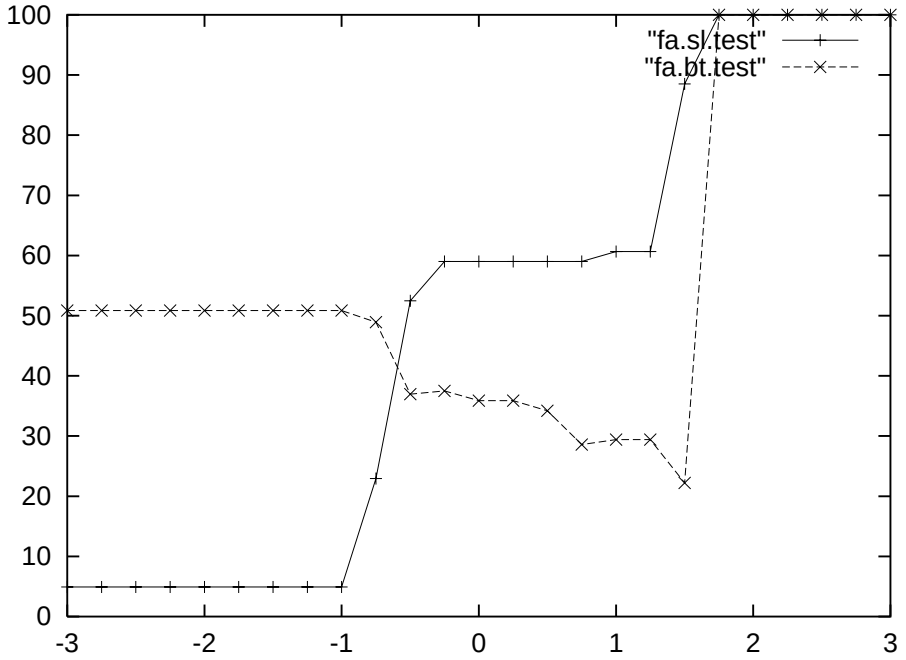


Figure 4: Évolution du bruit et du silence avec la moyenne fréquentielle pour les expressions

Expressions (conclusion)

- Aucune métrique ne peut donner un compromis bruit/silence qui soit satisfaisant (le bruit et le silence se croisent dans les alentours des 40 % pour chacun).
- Malheureusement, aucune combinaison ne donne de meilleurs résultats. On gardera donc l'entropie comme seule métrique de filtration.
- La fréquence est mise de côtée car, dans l'optique où l'on laisse le choix du seuil à l'utilisateur, une métrique la plus continue possible (qui prend le plus de valeurs) est souhaitable. L'utilisateur a ainsi plus de choix.
- De plus, on soupçonne l'entropie d'être encore plus efficace sur de grands corpus, contrairement à la fréquence.

Mots

- Un déroulement similaire est appliqué pour les mots. Dans ce cas-ci, le corpus de médecine est utilisé.
- La conclusion est la même: l'entropie est la meilleure métrique, et aucune combinaison n'est satisfaisante.
- De plus, la fréquence est encore une fois une bonne métrique, mais oubliée pour les mêmes raisons.

Inefficacité des métriques sur les termes de fréquence unitaire

Afin de justifier l'affirmation que les métriques ne fonctionnent pas sur les séquences de fréquence unitaire, voici un tableau démontrant ce fait:

N	20.00	40.00	60.00	80.00	100.00	120.00	140.00	160.00	180.00	200.00
f	0.00	2.50	5.00	7.50	8.00	9.17	8.57	7.50	10.56	9.50
l	0.00	5.00	5.00	3.75	5.00	5.00	5.71	6.25	6.67	6.50
d	0.00	5.00	5.00	3.75	5.00	5.00	5.71	6.25	6.67	6.50
dm	5.00	5.00	6.67	10.00	8.00	6.67	5.71	6.88	6.11	7.00
fag	0.00	2.50	5.00	7.50	8.00	9.17	8.57	7.50	10.56	9.50
mim	0.00	5.00	5.00	3.75	5.00	5.00	5.71	6.25	6.67	6.50
s	10.00	7.50	5.00	6.25	9.00	10.00	9.29	10.62	10.00	12.00
c	0.00	5.00	5.00	8.75	7.00	5.83	7.86	6.88	6.67	8.50
e	0.00	2.50	5.00	7.50	8.00	9.17	8.57	7.50	10.56	9.50
kuc	0.00	5.00	5.00	3.75	5.00	5.00	5.71	6.25	6.67	6.50
och	0.00	5.00	5.00	3.75	5.00	5.00	5.71	6.25	6.67	6.50
chi	0.00	5.00	5.00	3.75	5.00	5.00	5.71	6.25	6.67	6.50
smc	0.00	5.00	5.00	3.75	5.00	5.00	5.71	6.25	6.67	6.50
phi	0.00	5.00	5.00	3.75	5.00	5.00	5.71	6.25	6.67	6.50
mi	0.00	5.00	5.00	3.75	5.00	5.00	5.71	6.25	6.67	6.50
y	0.00	5.00	5.00	8.75	7.00	5.83	7.86	6.88	6.67	8.50

Table 3: Progression de la précision des métriques sur le corpus de l'eau pour les expressions de fréquence unitaire

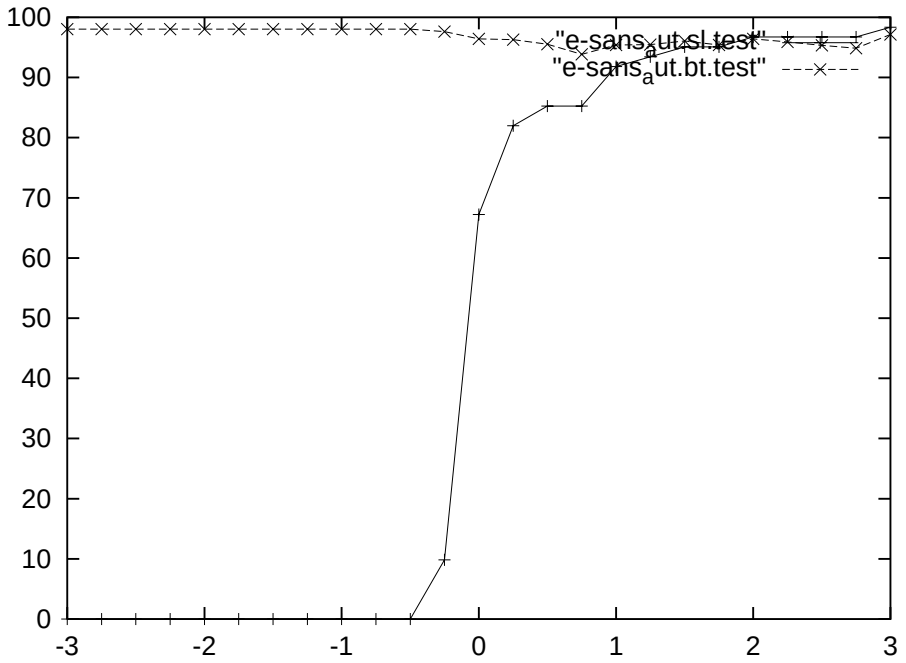
Évaluation des options

En plus des métriques, il y a des options du programme qui influe sur les termes extraits.

- Automate pour détecter les groupes nominaux
- Élimination des sous-séquences
- Fusion des variations morphologiques
- Fusion des variations terminologiques

Automate

Ce graphe montre très bien l'utilité d'un automate:



Élimination des sous-séquences

Éliminer les sous-séquences n'est peut-être pas toujours souhaitable

En effet, pour le corpus de l'eau, 10 % des termes de la liste de référence (expressions de fréquence 2 et plus) sont des sous-séquences. Par exemple:

- "eaux de pluie" apparaît toujours dans la séquence "les eaux de pluie et de la fonte des neiges"

Il n'est donc pas suggéré d'utiliser cette option.

Fusion des variations morphologiques

La fusion des variations morphologiques est aussi une bonne option:

Terme du corpus	Fréquence sans variation	Fréquence avec variation
analyse bactériologique	1	2
bactérie de type coliforme	1	2
champ d' épuration	7	8
contamination bactérienne	9	11
eau de pluie	1	3
eau de ruissellement	1	2
eau de surface	5	9
eau naturelle	2	3
eau souterraine	17	25
fosse septique	1	3
garantie d' eau	4	5
nappe d' eau	2	3
nappe souterraine	9	10
puits artésien	25	35
puits domestique	1	4
puits foré	4	6
puits municipal	1	2

Résultats finaux

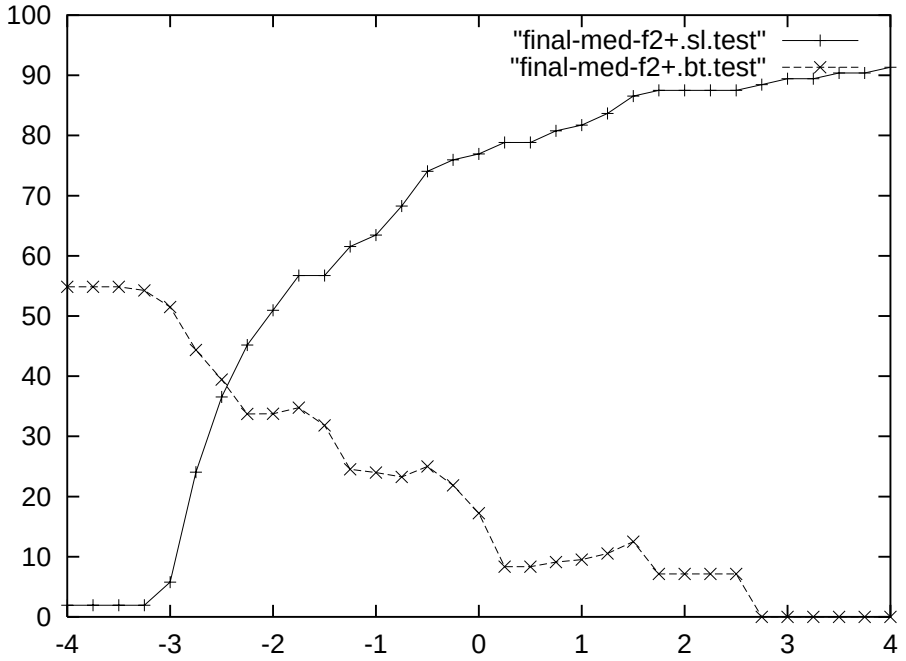


Figure 5: Évolution du bruit et du silence finale pour le corpus de médecine (mots et expressions de fréquence 2 et plus)

Résultats finaux (suite)

Mots	Expressions
est	corpus de l' eau
”	fréquence unitaire
“	liste de référence
corpus	termes extraits
mots	corpus de médecine
termes	ratio de vraisemblance
fréquence	affinités lexicales
métriques	corpus de médecine
liste	moyenne fréquentielle
entropie	fichier de configuration
expressions	métriques statistiques
$	expressions de fréquence
fait	
séquence	

Table 4: Comparaison des fréquences en considérant les variations morphologiques ou pas (termes singuliers)

Résultats finaux (suite)

Voici aussi une partie de la sortie CGI du programme:

...

Dans ce cas ci , la liste de référence contient 61 expressions de fréquence 2 et plus . Il semble donc que les métriques les plus efficaces soient la fréquence , le ratio de vraisemblance et l' entropie . La moyenne fréquentielle et le coefficient de Cosine réussissent passablement bien aussi . Pour trancher , on n' a qu' à observer le même tableau , mais pour le corpus

...

Ici , la liste de référence contient 39 expressions de fréquence 2 et plus . On peut observer que la fréquence , le ratio de vraisemblance et l' entropie sont toujours très efficaces . De plus , le coefficient de Dice modifié est aussi très bon . On ne peut cependant pas le retenir , car il ne donnait pas de bons résultats dans le corpus de l' eau .

...

Conclusions

Les métriques statistiques laissent croire qu'elles ne sont pas assez efficaces pour permettre un seuil unitaire d'extraction. Ceci peut être expliqué de deux façons:

- soit l'extraction est une activité trop subjective pour permettre un choix justifié pour chacun
- soit la sémantique des mots est une connaissance nécessaire à l'extraction terminologique

La détection des variations de termes est une autre voie, mais il est probable qu'elle ne saura pas combler totalement le vide de la sélection. Pour l'instant, on devra laisser le soin à l'utilisateur de faire le compromis entre silence et bruit. De plus, un terme peut souvent être de fréquence unitaire, un problème qu'on ne peut pas régler à l'aide de métriques statistiques.

Dans le cas des séquences à fréquence multiple, l'utilisation de l'entropie est donc la plus profitable. Il est intéressant de souligner que, comme pressenti dans les travaux de Béatrice Daille, la fréquence est une mesure plutôt efficace.

Conclusions (suite)

Ces conclusions sont conformes avec les résultats d'autres études. L'office de la langue français a testé plusieurs logiciels similaires:

Logiciels	Lexter	Nomino		System Quirk	TermFinder	Ztext
		UCN	UCN et UCNA			
Silence	22	12	7	59	39	78
Bruit	84	78	84	96	88	94

Ces résultats tiennent compte des séquences de fréquence quelconque.

On peut voir que les résultats ne sont pas très impressionnant. D'ailleurs, l'application la plus performante, Nomino, possède un module sémantique.

Voies futures

L'étude de l'extraction de termes est loin d'être terminée, et voici les chemins potentiellement avantageux:

- approfondissement de l'étude sémantique
- détection des variations terminologiques

Il existe cependant des termes qui n'apparaissent que sous des variations et que l'on devrait détecter. Par exemple:

- "carbonate de magnésium" dans "carbonate de calcium et de magnésium"
- "captage complet" dans "captage résidentiel complet".

Voies futures (suite)

Même si elle ne suffit pas à elle-même, une telle application peut déjà servir à construire un lexique spécialisé pour un domaine particulier. Un certain travail manuel devra par contre venir compléter l'extraction.

Dans un contexte bilingue, il peut aussi servir à constuire un dictionnaire bilingue spécialisé. Dans le cas où une personne possède un même texte dans deux langues différentes, la sortie de l'extracteur pour les deux corpus est alors étudiée par un modèle de traduction, afin de déterminer les associations traductives possibles.

Bibliography

G. F. Foster. 1991. Statistical lexical disambiguation. Master's thesis, School of Computer Science, McGill University.

Christian Jacquemin. 1997. *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus.* Ph.D. thesis, Université de Nante, Nantes.

Graham Russell. 1998. Identification of salient token sequences. Internal Report, RALI.

Autres programmes

Des applications complémentaires ont aussi été développées. Une d'entre elles est un détecteur d'abréviations. Il fonctionne pour français et anglais. Voici la sortie pour le corpus de médecine:

CD : clusters of differentiation
IFNa : Interferon alpha
IFNb : Interferon beta
IFNg : Interferon gamma
G-CSF : granulocyte colony stimulating factor
M-CSF : macrophage colony stimulating factor
GM-CSF : granulocyte-macrophage colony stimulating factor
IL : Interleukin
TNF : tumor necrosis factors
IL : Interleukins
TH2 : T helper
ELAM-1 : endothelial leucocyte adhesion molecule
ICAM-1 : intercellular adhesion molecule
VCAM-1 : vascular cell adhesion molecule
AIT : allergen immunotherapy
PBL : peripheral blood lymphocytes
AD : atopic dermatitis