Extracteur Terminologique Statistique
Stage du CRSNG, été 2002

Hugo Larochelle

2002

Table des matières

Table des figures

Liste des tableaux

1 Introduction

L'extraction terminologique, une activité linguistique courante, donne un défi sérieux à l'informatique, science de l'automatisation. En effet, on peut décrire un terme comme *une représentation littéraire d'un concept dans un domaine donné*[8]. L'acquisition de termes suppose donc celle du sens des mots, qui n'est qu'à un stade très peu développé dans le domaine de la linguistique informatique. Cependant, d'autres voix sont explorées afin de contourner ce problème. Il suffit de chercher des termes nous-même pour remarquer qu'il n'y a pas que la sémantique qui influence notre sélection terminologique dans un corpus donné. La fréquence ou la rareté d'un mot pèse aussi sur nos choix. Certaines métriques statistiques ont donc tenté d'utiliser ces échappatoires, et c'est leur efficacité qui sera traitée ici.

Pour ce faire, on a donc développé un extracteur terminologique, dont le fonctionnement sera d'abord décrit. Puis, l'exposé de l'évaluation des différents tests statistiques sera fait, celle-ci étant basée sur deux corpus mis à notre disposition, dont on a prélevé manuellement une liste de termes. Le premier est constitué de six textes de vulgarisation portant sur l'alimentation en eau et comporte 12492 mots. La sélection terminologique fut dirigée par des membres de l'Office de la langue française. Le second est un texte professionnel de médecine de 3296 mots, et l'extraction à la main fut produite par cinq membres du RALI/LLI de l'université de Montréal, dont moi-même. Non seulement on pourra vérifier le succès des différentes métriques à l'aide de ces corpus, mais certains résultats d'autres logiciels sont disponibles pour le premier, ce qui permettra aussi d'identifier les différentes avancées réalisées ailleurs. Finalement, certaines pistes seront abordées, afin de mieux orienter les futures recherches.

Il est à noter que tous les dossiers cités dans ce mémoire sont situés dans le répertoire `/home/contour/U/larocheh`. De plus, le présent mémoire est sous forme LaTeX et postscript dans le dossier `presentation/` de ce même répertoire.

2 Cheminement du corpus à analyser

Avant d'étudier l'analyse reliée à l'extraction terminologique, examinons les différentes étapes requises. Elles peuvent être divisées ainsi :

On suivra le modèle de cette ligne de commande Unix afin de structurer la description des étapes.

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

2.1 Corpus

Afin d'effectuer une étude précise des méthodes statistiques utilisées, on a à notre disposition deux corpus associés à une liste de termes extraits manuellement :

- Six textes de vulgarisation portant sur l'alimentation en eau totalisant 12492 mots. La liste de termes associée a été construite par des membres de l'Office de la langue française, dans le cadre du projet ATTRAIT (Atelier de travail informatisé du terminologue). Ce projet avait comme objectif d'évaluer des logiciels terminotiques répertoriés. Cette liste ne contient que des expressions de plus d'un mot. Les termes furent d'abord trouvés manuellement, puis certaines lacunes furent comblées à la vue de la sortie automatique des logiciels testés. Pour plus de renseignements, `http://www.rint.org/attrait/contexte.htm`.
- Un texte professionnel de médecine de 3296 mots. La liste de termes associée a été produite par cinq membres du RALI/LLI, dont moi-même. Chacun de ceux-ci a reçu un texte, dans lequel il a souligné les termes qu'il croyait pertinents. Un consensus fut atteint en sélectionnant les termes choisis par au moins trois personnes. Finalement, puisqu'en soulignant les termes, les membres n'indiquaient pas si une sous-séquence de ce terme était pertinente ou non, on a ajouté toutes les sous-séquences des termes sélectionnés si elles apparaissaient dans d'autres contextes dans le texte. Par exemple, le terme "white blood cell" contient le terme "cell", utilisé ailleurs. Ce dernier est donc retenu. La liste contient des mots et des expressions.

Après avoir fait la cueillette de terminologie manuelle, on se rend compte à quel point il est parfois difficile de faire une sélection judicieuse. Le choix n'est pas toujours évident, ce qui ouvre la porte à la subjectivité. D'ailleurs, le nombre de termes approuvés par personne varie de 99 à 343. De plus, comme le montre le tableau TAB. 1, l'unanimité n'est pas particulièrement répandue.

Nb de personnes	Nb de termes	Exemples
5	55	molecule, immune system, proinflammatory cytokines, ...
4	104	target cell, antibody, paracrine, ...
3	187	blood, patient, primary immune response, ...
2	269	chemotactic molecules, chills, therapeutic potentials, ...
1	427	organs, site, propagation of IgE, ...

TAB. 1 – Nb de termes sélectionnés en fonction du nombre de personnes minimum ayant choisi les termes

Comme on peut le voir, sur les 187 termes finalement sélectionnés, seulement 55 font l'unanimité, ce qui en laisse 49 qui sont appuyés par 4 personnes sur 5.

Même s'il est vrai qu'être membre du RALI/LLI ne fait pas de nous des terminologues, il est vraisemblable de croire que même les experts peuvent faire face à de telles situations. D'ailleurs, même à l'Office de la langue française, on a dû ajuster la sélection à la main après avoir observé les résultats fournis automatiquement.

Il est possible d'accéder aux deux corpus, dans le dossier `corpus/`. Ce répertoire en contient trois autres nommés `textes/`, `textes_tag/` et `selection_manuelle/`. Le premier contient les six textes sur la qualité de l'eau sous leur forme originale, en plus du fichier `corpus_eau.test` qui est une concaténation de ceux-ci. Il ne contient cependant pas le texte de médecine, mais il peut être récupéré à l'adresse suivante : `http://medic.med.uth.tmc.edu/hcprof/00000770.htm`. Le deuxième contient les deux corpus sous forme "tokenizée", "étiquetée" et "lemmatisée" (fichiers *.tag). Finalement, le dernier contient les différentes listes de termes extraits manuellement. Il y en a plusieurs, selon ce que l'on souhaite tester. Les noms contiennent les indications nécessaires pour les identifier. Si le nom contient le mot "med" ou "eau", c'est que la liste est associée respectivement au corpus de médecine ou à celui de l'eau, "expressions" et "mots" indique que la liste ne contient que des expressions (plus d'un mot) ou que des mots, et "f1" ou "f2+" spécifie que la liste ne contient que des termes de fréquence unitaire ou de fréquence 2 et plus. Finalement, l'ajout de l'abréviation "cf" précise que la liste a tenu compte des lemmes des mots ("citation forms"), en confondant entre autre les mots singuliers et pluriels.

Il est important de savoir que ces listes en question contiennent les termes sous leur forme apparaissant dans les corpus, et non pas sous leur forme neutre. La liste originale pour le corpus de l'eau est disponible à l'adresse suivante : `http://www.rint.org/attrait/Exprimentation/Extracteurs/annexeb.htm`. Pour ce qui est du corpus de médecine, les cinq listes de bases qui ont permis de converger vers la liste unique de référence sont disponibles dans le dossier `selection_manuelle/med/listesManuelles/`. Le fichier `notes.txt` associe à chaque membre du personnel du RALI/LLI la liste dont il est responsable.

2.2 Étiqueteur

Issu des travaux dans [5], l'étiqueteur utilisé permet de "tokenizer", i.e. séparer le texte en unités singulières ("tokens"), et d'étiqueter, i.e. ajouter la description grammaticale à ces unités, pour un corpus quelconque, en français ou en anglais. Par exemple, la phrase "Les enfants s'amusent dans le parc." devient :

Les	Dete-dart-ddef-masc-plur
enfants	NomC-masc-plur
s'	Pron-prfl-prea-genI-nomI-p3
amusent	Verb-IndPre-plur-p3
dans	Prep
le	Dete-dart-ddef-masc-sing
parc	NomC-masc-sing

.	Punc-pcst
{EOF}	

Comme on peut le voir, le mot et la grammaire sont séparés par des tabulations, détail important si l'on souhaite les séparer efficacement.

2.3 Lemmatiseur

À défaut d'avoir un lemmatiseur fonctionnant sous Linux, on a dû en programmer un. Un lexique pour le français et l'anglais étant disponible par l'entremise d'un lemmatiseur sous Solaris, on les a inclus dans des "Berkley Date Base"(BDB) distinctes. Un script Perl utilise alors une des BDB (selon la langue du texte) afin d'ajouter à chaque ligne du corpus étiquetté le caractère "/" suivi du lemme. L'appel du script se fait par la commande suivante :

addcf {db} [fichier]

où "db" est la base de donnée appropriée, et "fichier" est le nom du ficher à lemmatiser. Bien sûr, il est obligatoire de spécifier la base de donnée. Voici des exemples d'entrées dans un lexique sous forme texte :

industries	NomC	industrie
industriel	NomC	industriel
industriels	NomC	industriel

Il y en a deux de disponibles, soit "lexique.fr" et "lexique.en", pour le français et l'anglais, qui se trouvent dans le dossier addcitform/. On y trouve aussi les BDB "lexique.fr.data" et "lexique.en.data", de plus que l'application en question. Voici aussi un exemple de sortie du programme, appliquée à la même phrase que pour l'étiquetteur :

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

{EOF}

Il est à noter que la BDB a pour clefs tous les mots d'un lexique donné, mais sous forme minuscule.

2.4 Extracteur

L'architecture de l'extracteur de terminologie est relativement simple. Elle peut être décomposée ainsi :

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

2.4.1 Lecture du *corpus monde*

Avant de décrire cette étape, il est important de savoir ce qu'est le *corpus monde*. Idéologiquement, le *corpus monde* est l'ensemble de toute la littérature mondiale pour une langue donnée. Pratiquement, ce n'est qu'un recueil de corpus non spécialisés, donc qui vont dans plusieurs directions sémantiques. Dans ce cas-ci, on a utilisé l'ensemble du Hansard, recueil bilingue des débats parlementaires fédéraux canadiens. Préalablement, on a énuméré les divers mots de ce corpus et compté le nombre d'occurrences de chacun. À chaque mot est donc associée une fréquence dite "mondiale", et cette information est contenue dans les fichiers "corpusMonde.en" et "corpusMonde.fr", selon la langue. Les fichiers "corpusMonde.cf.en" et "corpusMonde.cf.fr" diffèrent dans le fait que les mots de même lemme sont confondus. Voici des exemples d'entrées que ces fichiers contiennent :

1360137	de
249766	je
2968	attitude
123	foresterie
5	criconscription

Ce sont ces fichiers qui sont lus par l'application. Cette information sera utilisée par deux des métriques exposées plus loin. Il est à noter que tous les mots sont sous forme minuscule.

2.4.2 Lecture du corpus à analyser

Ensuite, le corpus soumis en entrée est lu et stocké dans une liste, en y ajoutant, dans l'ordre, chacun des mots. Afin de sauver de l'espace mémoire, on utilise une référence pour les mots répétés dans le texte, i.e. les mots identiques alphabétiquement et grammaticalement. Aussi, les suffixes du texte sont reconstitués simultanément sous forme de StringModified, qui est en fait l'association d'un objet String avec l'index du suffixe dans le texte. Ceux-ci sont ensuite stockés dans une liste, afin de pouvoir créer le SFX et le LCP.

2.4.3 Création du SFX et du LCP

Les détails de la constitution du SFX et du LCP ne seront pas discutés ici. Pour plus de détails, veuillez vous référer à [11]. En bref, grâce à ces deux tableaux, il est possible d'obtenir rapidement la fréquence et les occurences de toute séquence apparaissant dans un corpus.

2.4.4 Recherche des séquences et assignation des scores

On a alors tous les éléments pour débuter l'analyse. Avant d'entrer dans les détails, il est important de faire la distinction entre mots et expressions. Ici, un mot peut comporter un espace. Par exemple, "de la" est considéré comme un seul mot. Un "token" serait probablement un terme plus juste mais, pour alléger le vocabulaire, on continuera à y référer comme un mot. Une expression est tout simplement en ensemble de plus d'un mot. La séparation entre mots et expressions est importante, puisque les métriques pour les expressions ne sont pas toutes transposables pour les mots. De plus, on parlera d'une séquence comme étant un mot ou une expression, et d'une sous-séquence comme étant une séquence apparaissant seulement à l'intérieur d'une autre séquence, bien sûr plus longue. On commence donc par trouver dans le corpus tous les mots et les expressions de fréquence 2 et plus. La raison est que les métriques disponibles ne semblent pas très efficaces dans le cas des termes à fréquence unitaire. Ceci sera discuté plus loin.

Ensuite, on crée une liste de mots et une liste d'expressions, puis on assigne, selon le cas, les valeurs aux différentes métriques appropriées. Différentes variables sont requises pour cela. Il y a bien sûr la fréquence **f** de la séquence dans le texte, à ne pas confondre avec la fréquence mondiale **F**, i.e. la fréquence d'un mot dans le *corpus monde*. Le fichier repésentant le *corpus monde* contient la valeur de **F** seulement pour les mots.

De plus, il y a les variables **a**, **b**, **c** et **d**, permettant de mesurer la liaison entre deux lemmes et définies par le tableau de contingence TAB. 2.

	B	¬B
A	a	b
$\neg A$	c	d

TAB. 2 – Tableau de contingence des critères d'association

On suppose qu'une séquence S a été divisée en sous séquence A et B ($S = AB$). Donc, par exemple, b est le nombre d'occurences de A sans être suivi de B dans le corpus. Puisque les expressions de 3 mots et plus peuvent être subdivisées de maintes façons, la valeur retenue parmi toutes les valeurs possibles pour une métrique spécifique sera la plus petite d'entre elles.

Voici donc la description de ces métriques :

Mots

- **Entropie (E)**

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

  où **s** est un séquence (dans ce cas-ci, un mot), **w** est un mot et **T** est le corpus. L'opérateur $|s|$ désigne ici la fréquence de **s** dans le corpus. Cette métrique est nulle lorsqu'une séquence apparaît devant et après les mêmes mots, et est maximale lorsqu'une séquence apparaît devant et après des mots toujours différents. Cette statistique est justifiée par le raisonnement suivant : une séquence pertinente devrait apparaître dans un nombre varié de contextes. L'entropie sera aussi utilisée pour les expressions. De plus, une modification de la fonction $h(x)$ sera apportée et expliquée plus loin. On utilisera plutôt $h(x) = -\log_2(x)$. Tirée de [12]. Voici un exemple :

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

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

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

où **M** est le *corpus monde*. Ici, l'opération $|T|$ désigne la taille du corpus (en mots). Cette métrique fut inspirée du Pouvoir de Résolution [10], applicable pour les affinités lexicales. Il est à noter que la fraction du logarithme représente l'approximation de la probabilité d'obtenir le mot **w** dans un texte quelconque. L'addition de **f** et de **F** permet d'éviter l'erreur causée par le logarithme de zéro, car **f** ne peut être nul. En bref, cette métrique favorise les mots rares. Voici un exemple :

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

Expressions

- **Ratio de vraisemblance (L)**

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

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. Tiré de [4].

- **Entropie (E)**

idem à l'entropie pour les mots.

- ***Chi*²**

$$Chi^2 = \frac{(a+b+c+d)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Cette mesure est bien connue dans le domaine de la statistique. Il est important de prendre note que dans le cas où **a**, **b**, **c** ou **d** est plus petit que 5, la formule suivante doit être utilisée :

$$Chi^2 = \frac{(|a+d-b-c|-N/2)^2}{(a+b)(c+d)(a+c)(b+d)}$$

- **Coefficient de Cosine (C)**

  $$C = \frac{a}{bc}$$

  Ce score peut créer des problèmes lorsque **b** ou **c** sont nuls. Dans un tel cas, on suggère de retourner le double de **a**, valeur qui ne peut être atteinte autrement.

- **Coefficient de Dice (D)**

  $$D = \frac{2a}{a+b+a+c}$$

  Cette mesure a été reportée comme ne surévaluant pas les séquences à faible fréquence. Tirée de [7].

- **Coefficient de Dice Modifié (DM)**

  $$DM = \log_2 \frac{2a^2}{a+b+a+c}$$

  Cette version du coefficient de Dice est supposée augmenter sa performance selon [9]. Tirée de [7].

- **Coefficient de Proximité Simple (SMC)**

  $$SMC = \frac{a+d}{a+b+c+d}$$

  Ce coefficient varie entre 0 et 1, et est symétrique entre A et B. Tiré de [3].

- **Coefficient de Kulczinsky(KUC)**

  $$KUC = \frac{a}{2}\left(\frac{1}{a+b} + \frac{1}{a+c}\right)$$

  Ce score varie aussi de 0 à 1, et lorsque A apparaît seulement avec B, il est supérieur à 0,5. Tiré de [3].

- **Coefficient d'Ochiai (OCH)**

$$OCH = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Ce score varie entre 0 et 1. Tiré de [3].

- **Coefficient de Fager et McGowan (FAG)**

$$FAG = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{a+b}}$$

Ce coefficient a 1 pour borne supérieure et ne l'atteint jamais. Il peut aussi être négatif. Tiré de [3].

- **Coefficient de Yule (Y)**

$$Y = \frac{ad-bc}{ad+bc}$$

Cette mesure varie entre -1 et 1 et est égale à 1 si A apparaît toujours avec B, ou l'inverse. Tirée de [3].

- **Coefficient de ϕ^2 (PHI)**

$$PHI = \frac{(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

Ce score a déjà été utilisé pour l'alignement de mots dans des phrases appariées par [6].

- **Information Mutuelle (MI)**

$$MI = \log_2\left(\frac{a}{(a+b)(a+c)}\right)$$

Ce score compare la probabilité de voir A et B apparaître ensemble avec celle de voir A et B séparémment. Elle fut décrite dans un contexte bilingue dans [1] et dans un contexte unilingue dans [2]. Tiré de [3].

- **Information Mutuelle modifié (MIM)**

$$MIM = \log_2\left(\frac{a^3}{(a+b)(a+c)}\right)$$

Cette version du score d'Information Mutuelle est issue des expérimentations dans [3]. Ayant essayé de la puissance 2 à 10, c'est le cube qui fut retenu, fournissant un compromis acceptable entre retenir les évènements rares et trop les négliger.

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

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

où **z** est un mot non-outil (i.e. un mot dont la fonction n'est pas purement syntaxique), **s** est la séquence à analyser, **f** est la fréquence de **z**, **F** est la fréquence mondiale de **z**, **T** est le corpus contenant **s** et **M** est le *corpus monde*. Ici, l'opération $|T|$ désigne la taille de T (en mots). Cette métrique fut aussi inspirée du Pouvoir de Résolution [10]. Bref, elle est la moyenne du score de comparaison avec le monde des mots non-outils de la séquence.

– **Moyenne fréquencielle (FA)**

$$FA = \frac{\sum_{z \in s} f}{|\{z|z \in s\}|}$$

Cette mesure est simplement la moyenne des fréquences des mots non-outils de la séquence. On a cru bon vérifier si cette mesure pouvait être utile.

De plus, la fréquence a aussi été testée comme outil statistique.

Bien sûr, ce ne sont pas toutes ces métriques qui sont utilisée lors de la filtration. Les étapes pour la sélection des bonnes métriques sont décrites dans la section Évaluation des Métriques. Après l'assignation des valeurs, la moyenne et l'écart type pour chacune des métriques sont calculés.

2.4.5 Filtration normale

Ayant calculé la moyenne et l'écart type pour chacune des métriques, il est alors possible de filtrer les bons termes à l'aide d'un seuil normal. Pour chaque mot et chaque expression, les valeurs aux métriques sont normalisées et comparées à un seuil, fixé par l'utilisateur. Les mots et expressions dépassant ou égalisant ce seuil sont alors imprimés à l'écran.

2.4.6 Options et paramètres

Voici la liste des options à la disposition de l'utilisateur :

xtt [–help][-l,e,chi,d,dm,c,smc,kuc,och,fag,y,phi,mi,mim,s,fa]
[-we,ws][-as,ami,amim][-v][-cgi][-fr][-la][-lao][-wo][-eo][-pos]
[-aut][-cf][-var][-shows][-t= {threshold}][file]

L'utilisateur peut choisir que seulement les mots soient affichés à l'écran (-wo), ou seulement les expressions (-eo). De plus, les expressions peuvent être classées selon plusieurs ordre. Par défaut, elles le sont par fréquence, mais il est possible de les ordonner selon chacune des métriques décrites plus haut. Même chose pour les mots. Par l'option -v (verbose), les différentes étapes de l'application sont imprimées lors de l'exécution. Il est aussi possible d'utiliser ce programme dans un script CGI, à l'aide de l'option -cgi. Celle-ci permet d'obtenir le code html d'une page affichant le texte, dans lequel les termes sélectionnés ont été soulignés. Puisque l'application est bilingue, il est important de spécifié la langue du corpus à analyser. Par défaut, c'est l'anglais, mais l'option -fr permet une analyse en français.

Certains paramètres permettent d'influer sur l'analyse elle-même. L'option -pos permet d'éliminer les sous-séquences à l'aide de la position de celles-ci, même s'elles pourraient être jugées pertinentes. Ensuite, -aut donne la possibilité de filtrer les séquences à l'aide d'un automate à état fini. Les automates permettent de s'assurer que les termes suivront un modèle grammatical bien spécifique. Puis, -cf autorise la confusion de mots ayant le même lemme (dans le cas où le corpus a été lemmatisé). L'option -var devait permettre d'utiliser les variations terminologiques, mais celle-ci n'est pas vraiment développée, entre autre parce que l'on n'a pas de moyen d'accéder au radical d'un mot. Finalement, à l'aide du paramètre -t, il est possible de déterminer le seuil de filtration des termes. Celui-ci doit être situé entre -4 et 4. Par défaut, aucune filtration n'est faite.

Une autre option fut initialement programmée, celle de l'affichage des affinités lexicales. Une affinité lexicale est un couple de mots qui apparaît souvent à distance raisonnable dans un corpus donné. Si le sujet n'est pas développé ici, c'est qu'aucune expérimentation n'y a été consacrée. Nous ne disposons donc pas de façon efficace de filtrer les affinités lexicales. Trois métriques sont cependant disponibles et permettent de les classer, soit le pouvoir de résolution (-as), l'information mutuelle (-ami) et l'information mutuelle modifiée (-amim). Par défaut, elles sont classées par ordre de fréquence. Pour ajouter l'affichage des affinités lexicales, utiliser l'option -la, pour obtenir un affichage exclusif, -lao.

Finalement, toute personne souhaitant obtenir la valeur aux différents tests pour les mots, expressions et affinités lexicales, n'a qu'à ajouter le paramètre -shows.

Pour plus d'information sur l'architecture de l'extracteur, voir la Javadoc située dans le répertoire `doc/`.

2.5 Problèmes ciblés

Mis à part le fonctionnement analytique de l'extracteur de termes, certaines difficultés ont été observées au niveau technique dans chacune des parties.

Corpus

- Les listes de termes de référence contiennent beaucoup de termes de fréquence unitaire, ce qui rend l'analyse statistique insuffisante puisque, comme souligné plus haut, les métriques existantes ne sont pas très efficaces dans ce cas. Il semble donc que d'autres voix devront être explorées.

Étiqueteur

- L'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.). Cependant, l'extracteur corrige ce problème, en les associant automatiquement à des ponctuations ;
- La segmentation du texte est quelque fois mal réalisée. Exemple :

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

- Lorsque l'étiqueteur fait face à un mot rare, il réussit plutôt mal à lui donner la bonne description grammaticale. Exemple :

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

- L'étiquetage ne peut être fait dans deux langues simultanément i.e., par exemple, que les mots en anglais d'un texte globalement en français seront mal étiquetés, et souvent associés à des noms propres.

Ces erreurs contribuent à diminuer l'efficacité de la détection des groupes nominaux (automate).

Lemmatiseur

- Les mots absents de la base de donnée fournie ne peuvent être lemmatisés ;
- Il serait plus utile d'ajouter le radical d'un mot, plutôt que le lemme. Ainsi, les mots apparaissant sous plusieurs formes auraient une plus grande fréquence, ce qui pourrait bien améliorer l'extraction. Aussi, si l'on souhaite éventuellement détecter les variations terminologiques du type X1 X2 $\rightarrow$ X1 of/de X2 (par exemple, puits résidentiel $\rightarrow$ puits de résidence), l'accès à la forme radicale est nécessaire.

Extracteur

- les phrases étant stockées indépendamment des mots, il est possible que le programme sature en mémoire trop rapidement dans le cas de gros corpus. Une approche utilisant des suites de mots (déjà en mémoire) plutôt que des phrases sous forme de String a été implantée, mais celle-ci doublait le temps d'exécution, ce qui n'est pas souhaité en Java, un langage déjà relativement lent. L'utilisation de Pat Trees pourrait régler ce problème ;
- Java n'étant pas très rapide, une implantation en C++ pourrait être souhaitable. Ceci règlerait les problèmes de rapidité apparaissant dans la tentative d'économie d'espace mémoriel, évoquée plus haut.

2.6 Autres Programmes

D'autres applications viennent compléter l'extracteur de terminologie. Dans le répertoire `writeAutomate/`, le programme wa permet d'écrire une classe Java décrivant un automate selon un fichier de configuration, qui contient en bref le nom des noeuds et les différents liens qui les unissent. Pour plus de détail sur la composition de ce fichier, voir la Javadoc du code (`writeAutomate/doc/`). La ligne de commande est la suivante :

wa [-n {name}][file]

où "name" est le nom à donner à la classe (par défaut, AutomateDefault). Le nom ne contient pas l'extension .java, celle-ci est ajoutée par le programme. Finalement, "file" est le nom du fichier de configuration. Deux sont actuellement disponibles, soit autFrancais.cfg et autEnglish.cfg, qui ont permis de créer les classes AutomateFrancais.java et AutomateEnglish.java.

De plus, le répertoire `abreviationSearch/` contient un petit logiciel permettant, à partir d'un fichier "tokenizé" et étiqueté, de chercher les abréviations qu'il contient ainsi que leur définition. Ceci peut-être utile dans les textes spécialisés, car les expressions complexes récurrentes sont souvent abrégées. Par exemple, pour le texte de médecine, l'application détecte les abréviations suivantes :

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

Bien sûr, l'application n'est pas parfaite et parfois ne réussit pas à détecter une abréviation parce que son initiateur a fait preuve de trop d'imagination (par exemple : TH : helper T cells), mais les textes en possédant plusieurs peuvent profiter de cet éclaircissement. Même si elle n'a pas été testée exhaustivement, elle devrait être relativement efficace, en français ou en anglais. L'appel est tout simple :

abSearch {cfgFile}[file]

où "file" est le nom du fichier texte "tokenizé" et étiqueté, et cfgFile est le fichier de configuration qui détermine les mots outils, i.e. les mots qui ne sont pas représentés par une lettre dans une abréviation. Par exemple, on définit les prépositions, déterminants, conjonctions et signes de ponctuation comme des mots outils par des lignes suivantes :

Prep
Punc
Dete
Con

Pour approfondir l'architecture de l'application, voir la Javadoc du programme dans le répertoire `abreviationSearch/doc/`.

Finalement, le répertoire `textes/` contient plusieurs petits textes (généralement moins de 1000 mots) en anglais, qui devaient initialement servir à tester l'extracteur par recherche d'information. Malheureusement, cette façon n'a pu être entamée. Ces textes n'ont donc plus vraiment d'utilité, mais sont toujours accessibles. Le dossier `textes/Token/` contient les mêmes textes sous forme "tokenizée" et étiquetée.

3 Évaluation des Métriques

On cherche maintenant à observer le travail fait par chacune des métriques décrites plus haut. L'objectif n'est pas nécessairement de trouver une seule métrique qui saura tout faire, mais peut-être d'en trouver qui pourront être combinées de façon efficace, ce qui semble plus réaliste. Il est à noter que l'automate, servant à détecter les groupes nominaux et utilisé pour l'étude des métriques n'est pas celui de la version finale, car celle-ci a été développée plus tard. La dernière version inclue plus de formes terminologiques.

Deux mesures seront utilisées de façon récurrente durant l'évaluation :

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. De plus, dans les représentations graphiques, le bruit est représenté par une ligne avec des points sous forme de "X". Pour le silence, ce sont des "+". L'axe des ordonnées est le pourcentage de silence ou de bruit, l'axe des abscisses est le seuil normal appliqué.

On débutera donc par les métriques pour les expressions, mais avant, certains changements du score de l'entropie doivent être effectués. En étudiant un peu la formule de ce test, on peut se rendre compte qu'il est possible de l'optimiser un peu. On a donc comparé l'originale (FIG. 1) avec l'utilisation d'une autre formule, $h(x) = -\log_2(x)$ (FIG. 2).

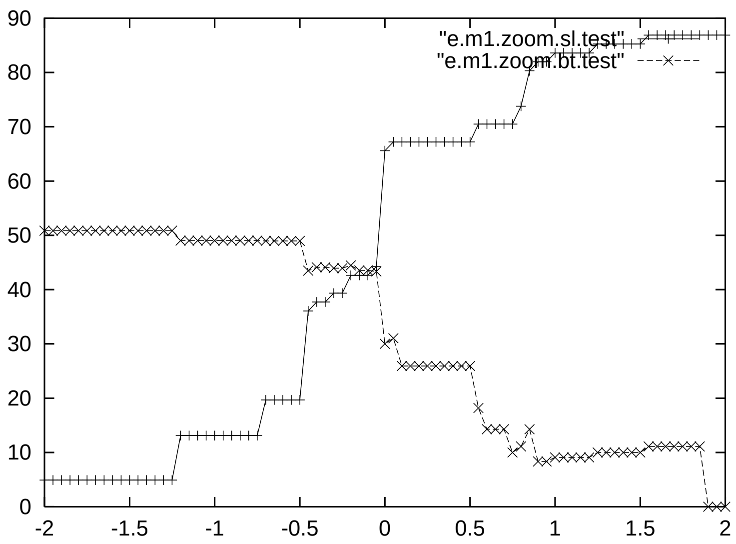

FIG. 1 – Évolution du bruit et du silence avec l'entropie originale

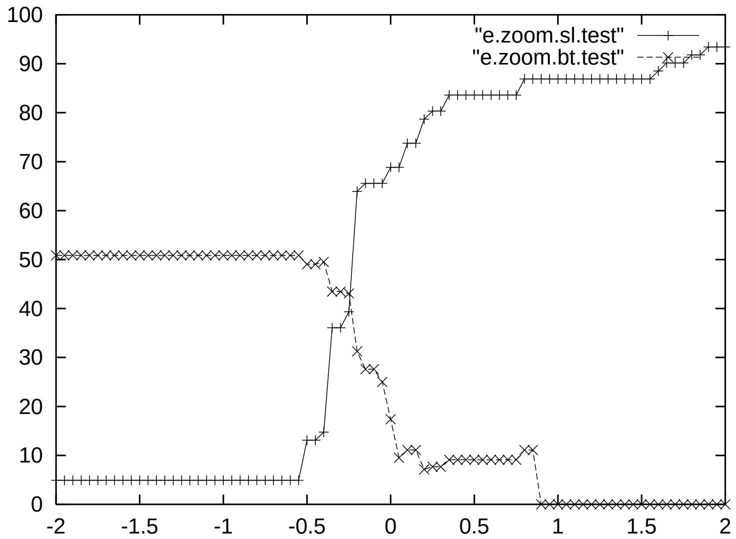

FIG. 2 – Évolution du bruit et du silence avec modification du score de l'entropie où $h(x) = -\log_2(x)$

Il est plutôt difficile de trancher. Cependant, en observant les valeurs plus précisément, on se rend rapidement compte que la modification est profitable. Quelques exemples sont exposés dans le tableau TAB. 3.

$h(x) = -x\log_2(x)$		$h(x) = -\log_2(x)$	
Silence	Bruit	Silence	Bruit
65.57	30.00	65.57	25.00
70.49	18.18	68.85	9.52
80.33	14.29	78.69	7.14

TAB. 3 – Comparaison des scores d'entropie avec $h(x) = -x\log_2(x)$ et $h(x) = -\log_2(x)$

Il est facile d'expliquer ces résultats si l'on observe le comportement des fonctions $-x\log_2(x)$ et $-\log_2(x)$ pour des petites valeurs de x. La deuxième augmente toujours lorsque x diminue, alors que la première augmente puis diminue (voir FIG. 3).

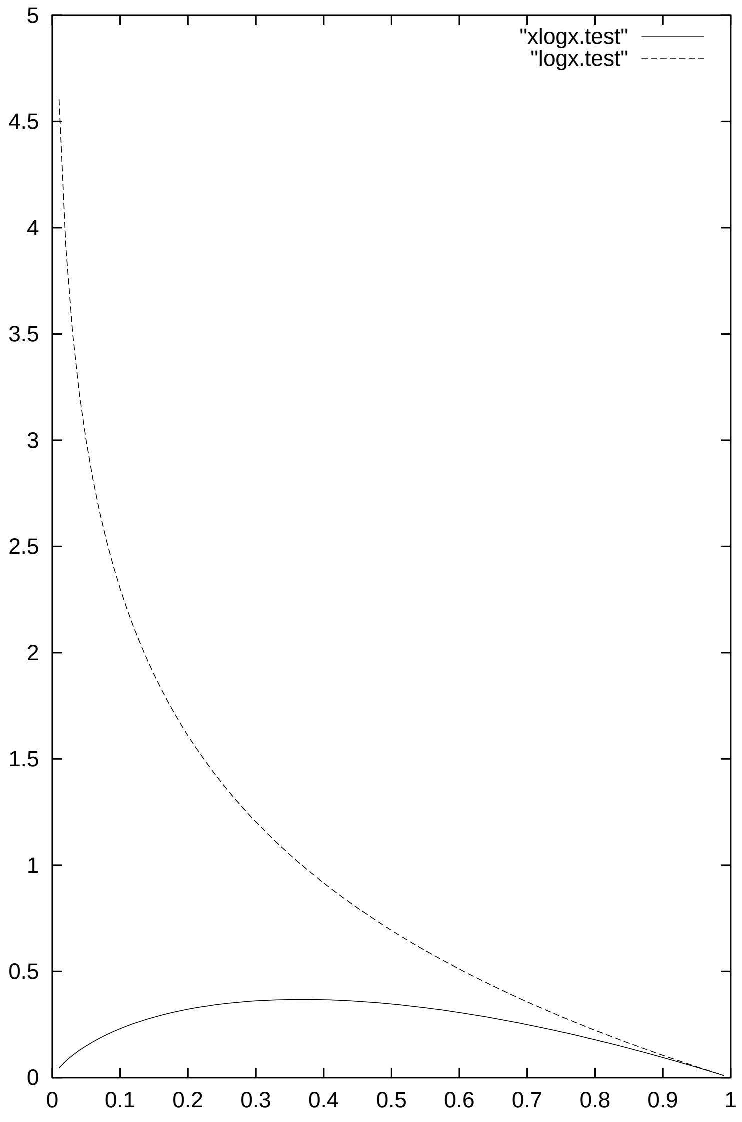

FIG. 3 – Comparaison des fonctions $h(x) = -x\log_2(x)$ et $h(x) = -\log_2(x)$ pour $x < 1$

3.1 Expressions

Puisque l'on dispose de deux corpus associés à une liste de référence contenant des expressions, on doit choisir laquelle aura priorité sur l'autre. On a donc décidé que les différentes évaluations se feront d'abord sur le corpus de l'eau, étant issu du travail de professionnels, i.e. les gens de l'Office de la langue française.

Pour débuter, on a ordonné la liste des expressions du corpus de l'eau par rapport à chacune des métriques de façon décroissante. Ces séquences sont toutes des groupes nominaux de fréquence supérieure ou égale à 2 . On a ensuite calculé le pourcentage des **N** premières pour chaque métrique se trouvant dans la liste de référence. Les résultats sont affichés dans le TAB. 4.

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

TAB. 4 – Progression de la précision des métriques sur le corpus de l'eau pour les expressions

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, pour le corpus de médecine (voir TAB. 5).

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

TAB. 5 – Progression de la précision des métriques sur le corpus de médecine pour les expressions

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 aussi 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. Finalement, la moyenne fréquentielle continue à réussir passeblement bien, mais le coefficient de Cosine, lui, est plutôt inefficace. On rejète donc Cosine, mais on gardera la moyenne fréquenctielle. Il est à noter que le score de comparaison avec le monde est beaucoup plus utile dans le corpus de médecine que dans le corpus de l'eau. On croit que ceci peut être expliqué par le fait que la médecine nécessite un vocabulaire beaucoup plus spécialisé que celui du domaine de l'aqueduc. Des mots comme “eau”, “puits” et “potable” sont relativement connus et utilisés fréquemment.

On a donc sélectionné 4 métriques parmi les 17 possibles, soit la fréquence, l’entropie, le ratio de vraisemblance et la moyenne fréquentielle. Pour mieux étudier le travail de chacune, on a tracé un graphique affichant l’évolution du bruit et du silence d’une métrique, en fonction du seuil normal. La fréquence est présentée par FIG. 4, l’entropie par FIG. 5, le ratio de vraisemblance par FIG. 6 et la moyenne fréquentielle par FIG. 7. Ces graphes ont été construit en utilisant le corpus de l’eau.

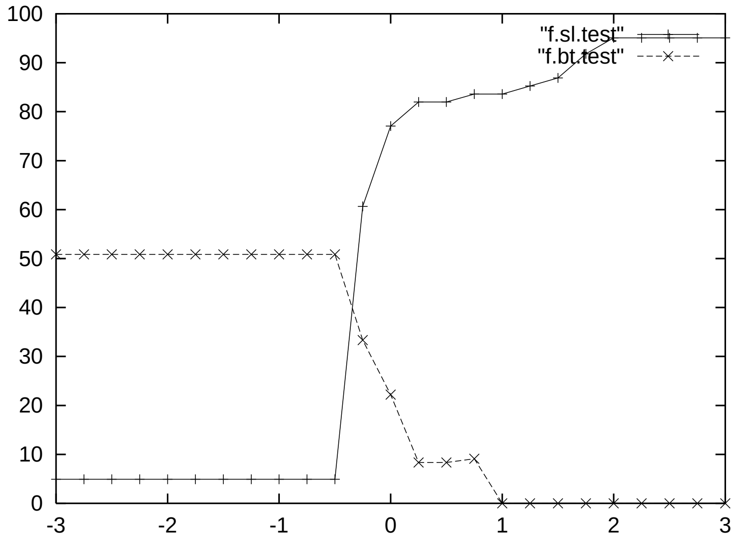

FIG. 4 – Évolution du bruit et du silence avec la fréquence pour les expressions

On fait alors face à un problème. Il semble qu’aucune métrique puisse donner un compromis bruit/silence qui soit satisfaisant (le bruit et le silence se croise dans les alentours des 40 % pour chacun). Pour l’instant, la seule possibilité est de laisser à l’utilisateur le choix de ce compromis. On voit que l’entropie et la fréquence permettent ceci, car elles peuvent causer beaucoup de bruit et peu de silence, peu de silence et beaucoup de bruit, ou un bruit et un silence moyen. Le ratio de vraisemblance peut lui aussi être utilisé ainsi, mais il semble avoir de la difficulté à diminuer le bruit. La moyenne fréquentielle ne réussit tout simplement pas à le faire. On la mettra donc de côté.

En combinant certaines d’entre elles, on arrivera peut-être à diminuer le

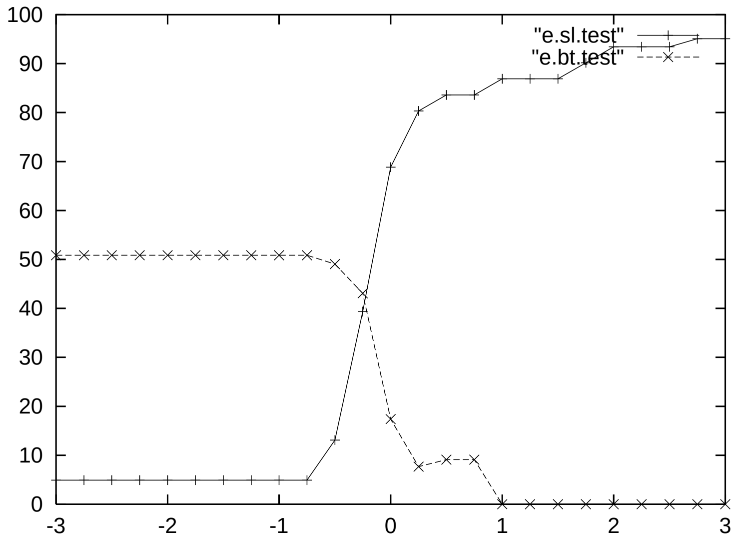

FIG. 5 – Évolution du bruit et du silence avec l'entropie pour les expressions

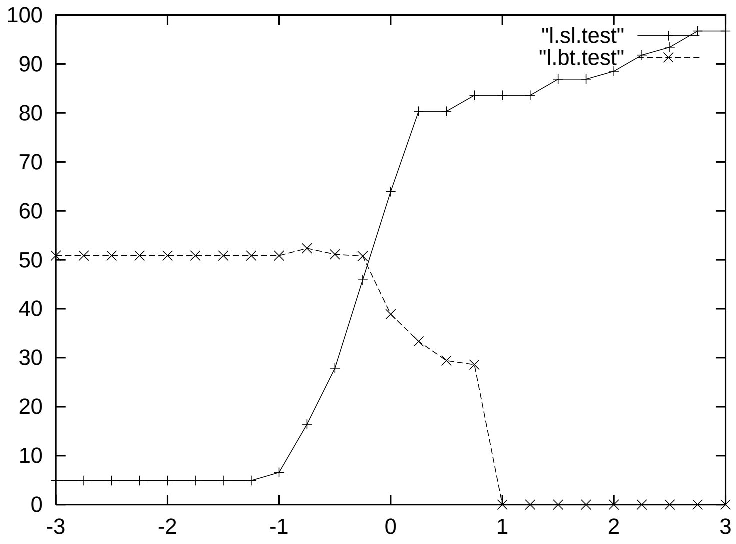

FIG. 6 – Évolution du bruit et du silence avec le ratio de vraisemblance pour les expressions

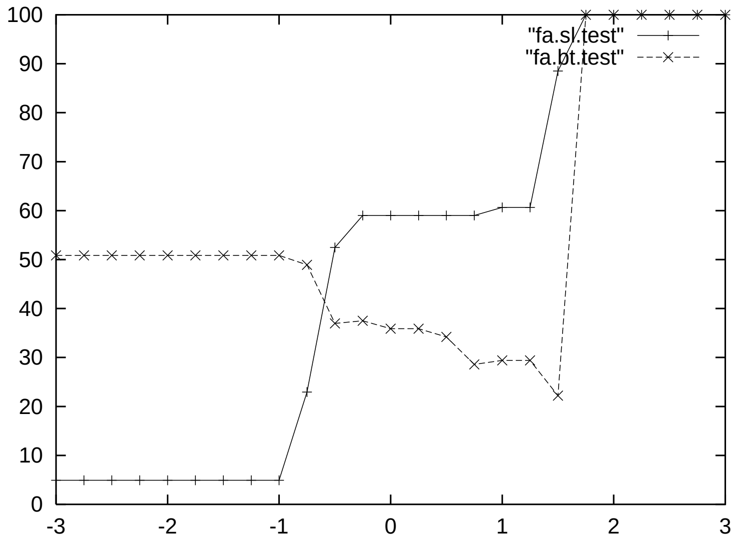

FIG. 7 – Évolution du bruit et du silence avec la moyenne fréquentielle pour les expressions

bruit et le silence simultanément. Pour l'instant, il semble que l'entropie et la fréquence soit les meilleurs tests. On priorisera tout de même l'entropie car, dans l'optique où l'on pourrait laisser le choix du seuil à l'utilisateur, on aura besoin d'une métrique la plus continue possible, afin de laisser un choix plus vaste de filtration. La fréquence, ne prenant que des valeurs entières, est limitée dans ce sens. De plus, on soupçonne l'entropie d'être plus efficace avec de plus grands corpus.

Afin de savoir quelles métriques peuvent être associées, on a constitué différents tableaux comparant le travail de celles-ci. Le premier (TAB. 6) compare les 30 premiers termes extraits automatiquement et se trouvant dans la liste de référence (précision) et donne le pourcentage en commun. Le second (TAB. 7), compare les 30 premiers termes extraits automatiquement ne se trouvant pas dans la liste de référence (bruit) et donne le pourcentage en commun. Finalement, le dernier (TAB. 8) fait la soustraction entre la précision et le bruit. Le corpus de l'eau fut utilisé.

Il existe deux façons de combiner des métriques : de façon conjonctive ou disjonctive. La conjection consiste à retenir un terme s'il a des valeurs aux métriques en question supérieures au seuil de chacune d'elles. Par la disjonction, on ne conserve que les termes qui ont au moins une valeur de métrique

	f	l	e
f	100	60	87
l	60	100	67
e	87	67	100

Tab. 6 – Précision commune entre les métriques pour les expressions (pourcentage)

	f	l	e
f	100	53	73
l	53	100	43
e	73	43	100

Tab. 7 – Bruit commun entre les métriques pour les expressions (pourcentage)

	f	l	e
f	0	7	14
l	7	0	24
e	14	24	0

Tab. 8 – Différence entre la précision et le bruit pour les expressions (pourcentage) supérieure à son seuil respectif. Les tests qui profiteront de la conjection sont ceux qui ont une plus grande précision. Pour la disjonction, c'est l'inverse, i.e. que ce doit être le bruit commun qui soit plus grand. Comme toute les valeurs du TAB. 8 sont positives, on n'utilisera que la conjonction. Il est à noter que, pour que la conjonction soit efficace, il ne faut pas seulement que la précision commune soit plus grande que le bruit commun, mais aussi qu'elle soit élevée.

On doit maintenant déterminer quels seront les tests que l'on combinera. L'entropie et le ratio de vraisemblance devraient être un bon choix. En effet, ils ont une précision relativement semblable (67 %) et un bruit plus différent (47 %). Ensuite, on tentera de jumeler la fréquence avec l'entropie selon un raisonnement semblable.

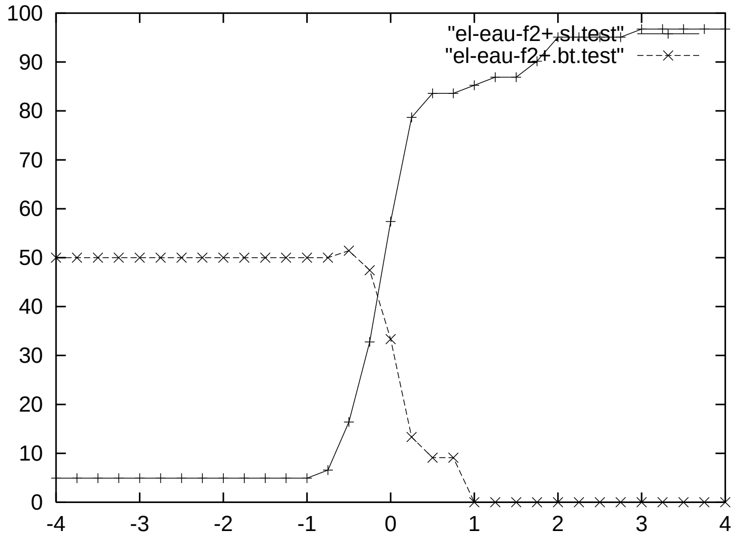

FIG. 8 – Évolution du bruit et du silence avec la combinaison de l'entropie et du ratio de vraisemblance pour les expressions

Malheureusement, aucune combinaison ne donne de meilleurs résultats. Les graphes sont d'ailleurs très similaires à celui de l'entropie seule. On gardera donc l'entropie comme seule métrique de filtration.

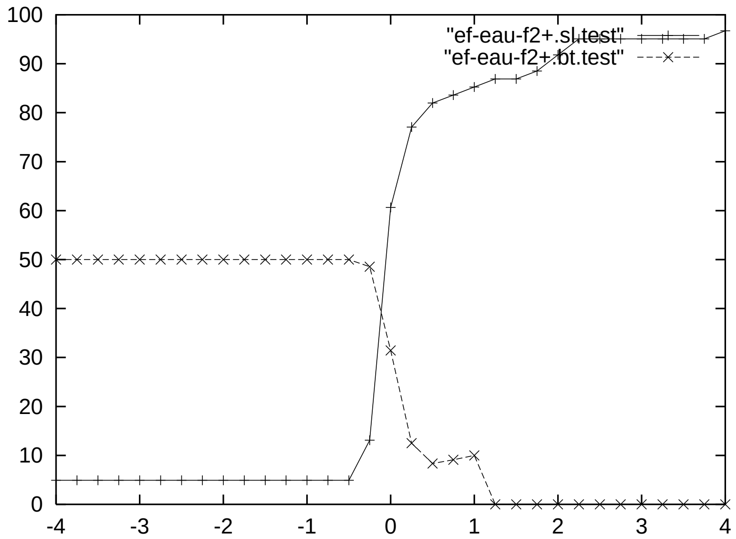

FIG. 9 – Évolution du bruit et du silence avec la combinaison de l'entropie et de la fréquence pour les expressions

3.2 Mots

Le déroulement est similaire pour les mots. Cependant, on ne dispose que du corpus de médecine pour faire nos tests. La liste de référence contient 64 termes (mots), qui sont bien sûr de fréquence 2 et plus. Le tableau TAB. 9 illustre la progression de la précision des métriques disponibles.

N	10.00	20.00	30.00	40.00	50.00	60.00	70.00	80.00	90.00	100.00
f	80.00	80.00	66.67	57.50	56.00	55.00	54.29	55.00	52.22	48.00
e	80.00	75.00	70.00	70.00	56.00	58.33	55.71	53.75	50.00	50.00
s	90.00	80.00	73.33	67.50	64.00	65.00	60.00	57.50	54.44	53.00

TAB. 9 – Progression de la précision des métriques sur le corpus de médecine pour les mots

On voit que le score de comparaison avec le monde est le plus efficace. Ceci est expliqué par le fait qu'un texte de médecine emploie un vocabulaire très spécialisé. Aussi, la fréquence et l'entropie sont relativement semblables, quoi que la dernière soit plus efficace aux alentours de 64 (soit $N = 60$ et $N = 70$). Pour l'instant, aucune sont éliminées. Les graphes de la progression du bruit et du silence est disponible pour la fréquence (FIG. 10), l'entropie (FIG. 11) et le score de vraisemblance (FIG. 12).

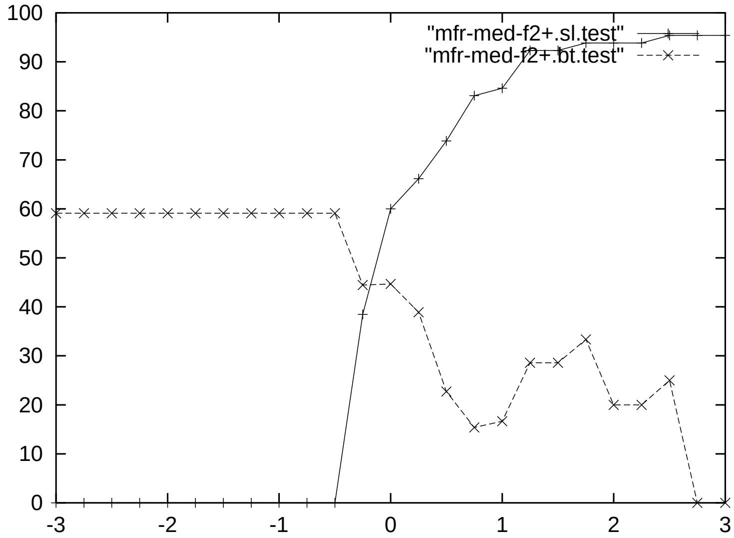

FIG. 10 – Évolution du bruit et du silence avec la fréquence pour les mots

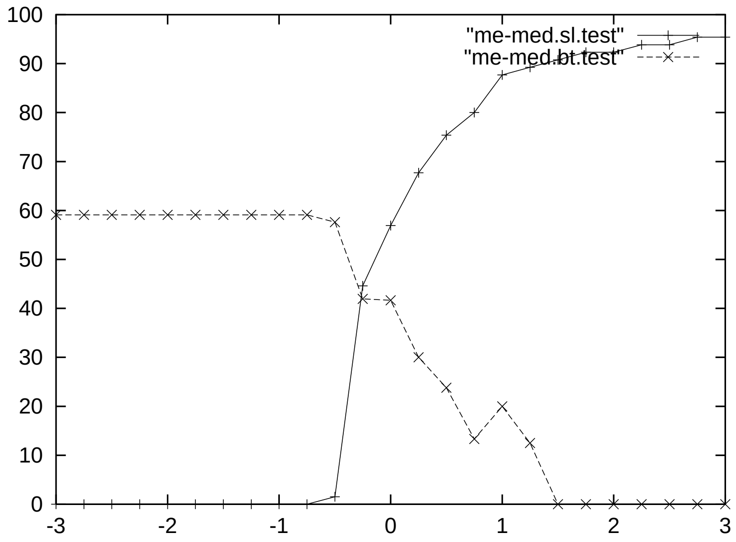

Fig. 11 – Évolution du bruit et du silence avec l'entropie pour les mots

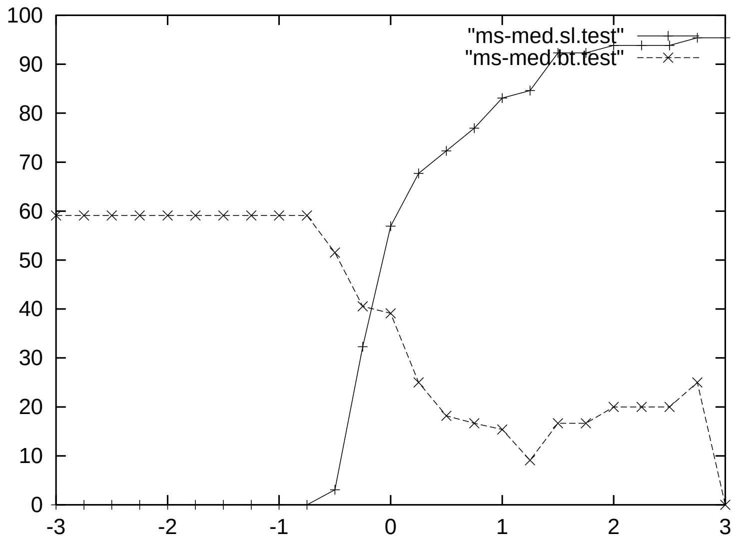

Fig. 12 – Évolution du bruit et du silence avec le score de comparaison avec le monde pour les mots

Encore une fois, il n'y a pas de compromis bruit/silence avantageux qui s'impose. On tentera donc, une fois de plus, de jumeler des métriques, en étudiant les tableaux de la précision commune (TAB. 10), du bruit commun (TAB. 11) et de la différence entre ces deux mesures (TAB. 12). Le tableau de la précision commune a comparer les 32 premiers termes extraits, et celui du bruit commun, 30.

	f	e	s
f	100	100	91
e	100	100	91
s	91	91	100

TAB. 10 – Précision commune entre les métriques pour les mots(pourcentage)

	f	e	s
f	100	83	90
e	83	100	77
s	90	77	100

TAB. 11 – Bruit commun entre les métriques pour les mots (pourcentage)

	f	e	s
f	0	17	1
e	17	0	14
s	1	14	0

TAB. 12 – Différence entre la précision et le bruit pour les mots (pourcentage)

À la vue de ces résultats, on essaie d'associer la fréquence avec l'entropie (FIG. 13), l'entropie avec le score de comparaison avec le monde (FIG. 14), ainsi que toutes les trois métriques (FIG. 15).

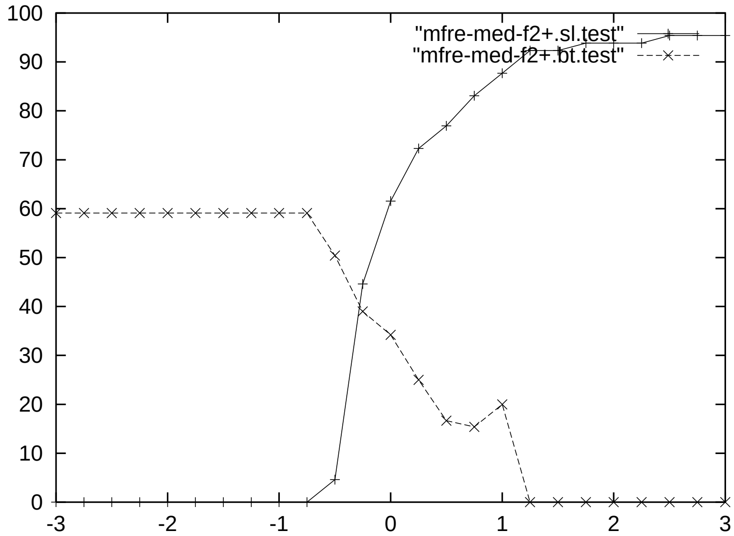

FIG. 13 – Évolution du bruit et du silence avec la combinaison de la fréquence et de l'entropie pour les mots

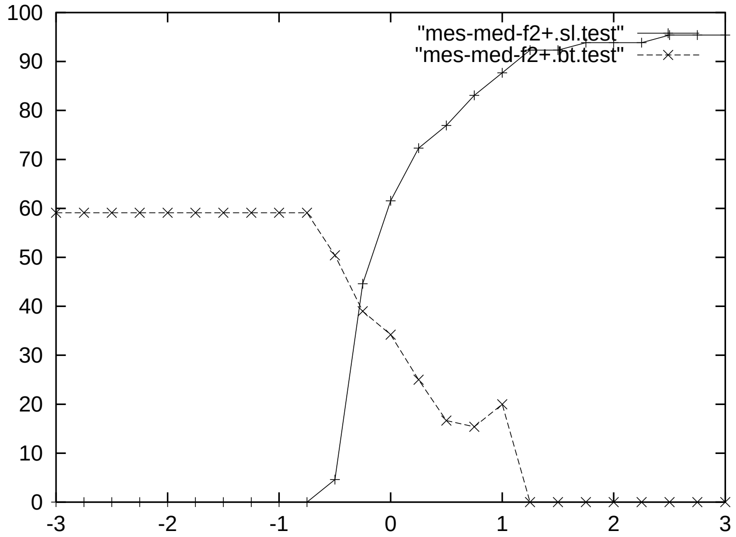

Fig. 14 – Évolution du bruit et du silence avec la combinaison de l'entropie et du score de comparaison avece le monde pour les mots

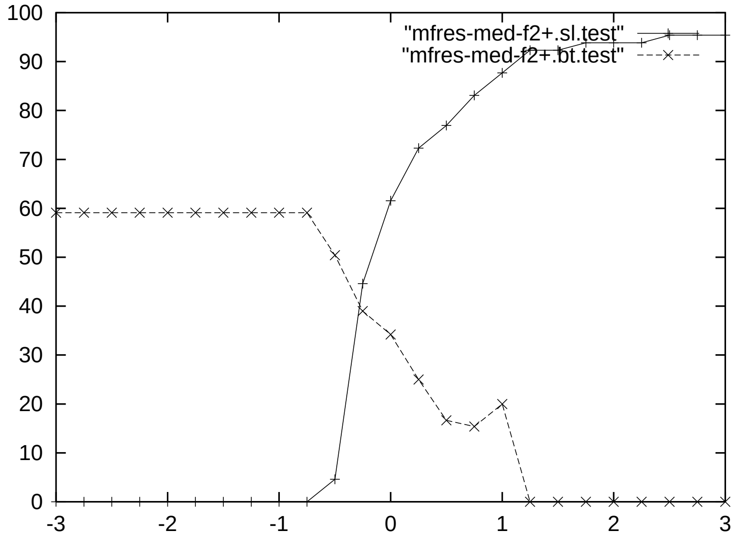

Fig. 15 – Évolution du bruit et du silence avec la combinaison de toutes les métriques pour les mots

On peut voir que les résultats sont les mêmes, quelle que soit la combinaison, et que l'amélioration est presque nulle par rapport à l'entropie seule. L'entropie est, encore une fois, choisie comme test unique.

3.3 Note sur les termes de fréquence 1

Comme indiqué plus haut, l'analyse des séquences ne se fait que sur celles qui apparaissent plus d'une fois dans le corpus. La raison est simplement que les métriques utilisées fonctionnent mal sur les séquences de fréquence unitaire. Le tableau TAB 13 montre bien ce fait. Le corpus de l'eau fut utilisé pour composer ce tableau.

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

TAB. 13 – Progression de la précision des métriques sur le corpus de l'eau pour les expressions de fréquence unitaire

Une des options possibles afin de contourner ce problème serait de développer une façon d'étudier la sémantique des mots de façon automatique. Cette intuition est appuyée par les résultats du projet ATTRAIT (voir TAB. 14).

En effet, le logiciel ayant le mieux réussit, Nomino, possède un module sémantique. Aussi, le bruit est élevé pour tous les logiciels, ce qui démontre à

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

TAB. 14 – Résultats d'autres logiciels sur le corpus de l'eau (en pourcentage)

quel point il reste beaucoup de travail à faire. Finalement, il est à noter que ces résultats ont été recueillis pour les expressions de fréquence quelconque. On ne peut donc pas les comparer avec ceux de l'extracteur décrit dans ce mémoire, puisqu'il est impuissant avec les séquences de fréquence unitaire.

Plusieurs autres tableaux et graphiques se trouvent dans les répertoires `tableaux/` et `graphiques/`, qui sont divisés en deux afin de séparer les résultats pour les mots et les expressions. La nomenclature des fichiers est la même que pour les listes de référence. Aussi, les répertoires `graphiques/expressions/filtresConj/` et `graphiques/mots/filtresConj/` contiennent les graphes pour les combinaisons conjonctives. Finalement, les graphes portant sur la modification de la formule de l'entropie sont situés dans le dossier `graphiques/entChange/`.

4 Évaluation des options

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

4.1 Automate

Il n'y a pas de façon infaillible de trouver le meilleur automate, pour l'anglais et pour le français. On a seulement regarder les deux listes de référence et essayer de rassembler toutes les formes s'y trouvant. La graphe FIG. 16, montrant la progression de la filtration avec l'entropie sans l'utilisation d'un automate (pour le corpus de l'eau), montre à quel point cette option est essentielle.

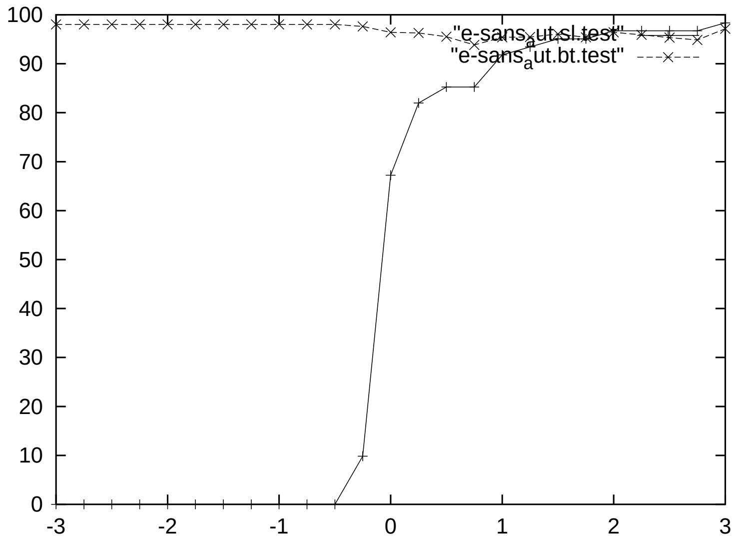

FIG. 16 – Évolution du bruit et du silence avec l'entropie sans l'utilisation d'un automate pour les expressions.

4.2 Élimination des sous-séquences

Pour éliminer les sous-séquences, on a utilisé la position des séquences. Cependant, certaines sous-séquences peuvent être des termes pertinents. En effet, pour le corpus de l'eau, 7 % des termes de la liste de référence (expressions de fréquence 2 et plus) sont des sous-séquences.

Par exemple, "système d'alimentation" apparaît toujours dans la séquence "système d'alimentation en eau".

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

4.3 Fusion des variations morphologiques

Les variations morphologiques sont très fréquentes dans un texte, ce qui rend essentiel leur prise en considération. Le tableau TAB. 15 donne un aperçu de l'importance de la confusion de ces variations. Il ne contient que des termes singuliers.

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

Tab. 15 – Comparaison des fréquences en considérant les variations morphologiques ou pas (termes singuliers)

4.4 Fusion des variations terminologiques

Même si l'option est disponible, elle n'est pas profondément développée. À vrai dire, elle ne fonctionne que pour les termes de longueur 2. Celui qui souhaiterait continuer le travail devrait consulter [8]. On y explique comment détecter les variations terminologiques. Ceci pourrait peut-être aider à inclure des termes de fréquence unitaire dans la recherche statistique. Cependant, il est plutôt improbable de croire que cela sera totalement satisfaisant. Voici pourquoi.

Il existe deux façons d'utiliser les variations terminologiques. La première est la suivante : on trouve préalablement les termes pertinents du corpus comme à l'habitude, puis on recherche les variations de ces termes et décidons si elles sont aussi pertinentes. Par contre, les variations alors trouvées seront de longueur 3 et plus et, pourtant, la liste de référence des expressions de fréquence unitaire en contient environ 40 % de longueur 2. Ajoutons à ça les termes qui n'ont pas de variations (puisqu'il y en a), et il semble de plus en plus que l'on travaille pour rien. Sinon, une deuxième méthode peut être utilisée. Il suffit de recueillir le plus de séquences à fréquence unitaire possible, de les inclure avec celles de fréquence multiple, et de regrouper celles qui sont dans une même famille de variation. On assigne alors à chaque groupe un score, afin de les comparer avec les séquences non regroupées. Le problème qui se pose maintenant est celui de s'assurer que la façon d'assigner ces scores pour chacune des métriques à une famille ne la surestime pas et ne la sous-évalue pas. De toute façon, en observant quelque peu le corpus de l'eau, on se rend compte que la plupart des termes à fréquence unitaire n'apparaissent que sous une seule forme.

Il existe cependant des termes qui n'apparaissent que sous des variations et que l'on devrait détecter, et c'est probablement l'utilisation la plus pratique que puisse avoir l'étude des variantes terminologiques. Par exemple, "carbonate de calcium et de magnésium" cache le terme "carbonate de magnésium", et "captage complet" est voilé par la séquence "captage résidentiel complet". Malheureusement, dans le corpus de l'eau, ces termes n'apparaissent qu'une seule fois, ce qui renvoie au même problème d'évaluation des termes à fréquence unitaire.

5 Résultats finaux

En appliquant la dernière version de l'automate et en utilisant seulement l'entropie, on obtient donc les graphes FIB. 17 pour le corpus de l'eau et FIB. 18 pour le corpus de médecine. Le dernier montre le bruit et silence pour l'extraction des mots **et** des expressions pertinentes.

Le croisement entre le bruit et le silence se fait dans les deux cas aux alentours de $t = -2.25$.

Afin d'avoir une idée de la sortie du programme, les termes extraits du mémoire que vous lisez en ce moment avec $t = 0$ sont affichés dans le tableau TAB. 16.

Mots	Expressions
est	corpus de l' eau
séquence	fréquence unitaire
fait	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

TAB. 16 – Sortie de l'extracteur pour le mémoire

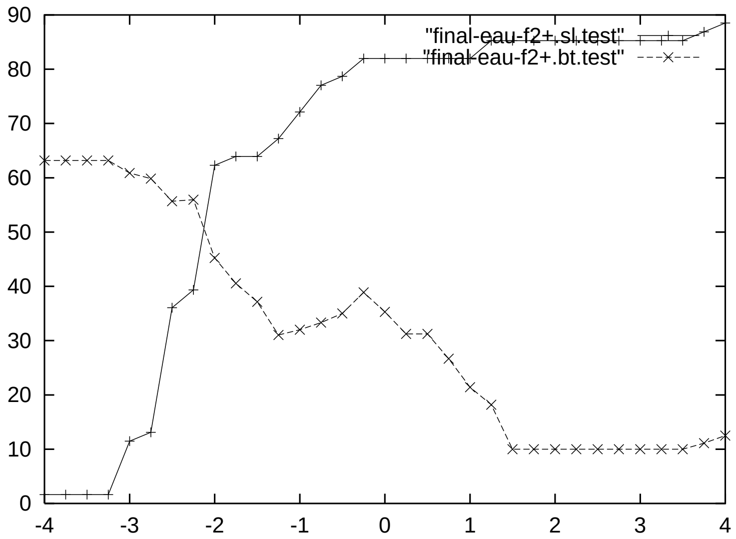

FIG. 17 – Évolution du bruit et du silence finale pour le corpus de l'eau (expressions de fréquence 2 et plus)

De plus, voici un aperçu de la sortie CGI, toujours pour le mémoire :

...

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 .

...

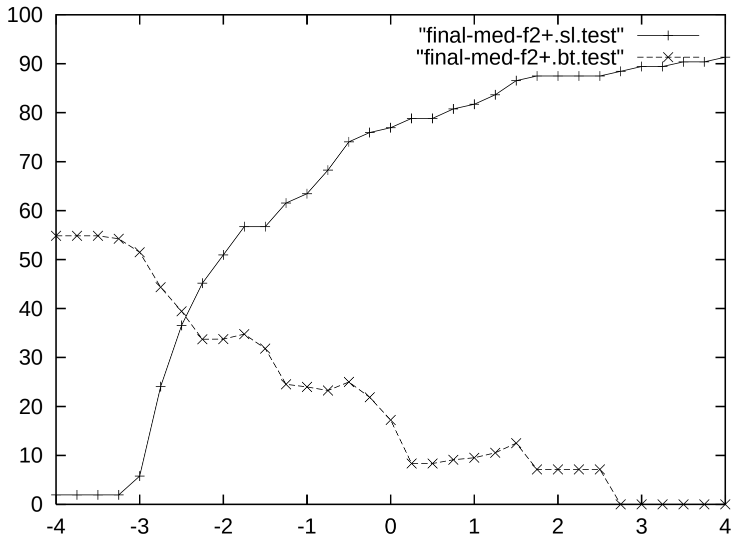

FIG. 18 – Évolution du bruit et du silence finale pour le corpus de médecine (mots et expressions de fréquence 2 et plus)

6 Conclusion

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

– soit l'extraction est une activité trop subjective pour permettre un choix justifié pour chacun. En effet, le manque d'objectivité a été évoqué ;
– 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 auparavant (voir [3]), la fréquence est une mesure plutôt efficace.

7 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. Effectivement, en plus de possiblement combler les lacunes de l'analyse statistique, elle est le seul moyen à notre disposition dans le cas des séquences à fréquence unitaire. Pour se faire, l'utilisation de WordNet est probablement la plus appropriée ;
- détection des variations sémantiques. La détection des variations pourra peut-être combler un certain vide, mais elle ne devrait pas être suffisante à elle-même, en particulier dans les petits corpus ;
- utilisation des affinités lexicales. Le potentiel de ce concept n'a toujours pas été totalement développé. La classification de textes à l'aide d'affinités lexicales a cependant déjà été réalisée. Ce serait peut-être une autre façon d'apprivoiser la sens dans l'extraction terminologique.

Même si elle ne suffit pas à elle-même, l'application exposée 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, elle 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.

Références

[1] Peter F. Brown, John Cocke, Stephen A Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, Robert L. Mercer, and Paul S. Roosin. A statistical approach to language translation. In *Proceeding of the 12th International Conference on Computational Linguistic (Coling-88)*, Budapest, Hungary, August 1988.

[2] Kenneth Ward Church and Patricia Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1) :22–29, March 1990.

[3] Béatrice Daille. *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7, 1994.

[4] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 1993.

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

[6] William A. Gale and Kenneth W. Church. Concordances for parallel texts. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research, Using Corpora*, pages 40–62, Oxford, U.K., 1991.

[7] Toru Hisamitsu and Yoshiki Niwa. Extracting useful terms from parenthetical expressions by combining simple rules and statistical measures : A comparative evaluation of bigram statistics. In D. Bourigault, C. Jacquemin, and M-C. L'Homme, editors, *Recent Advances in Computational Terminology*, pages 209–224. John Benjamins Publishing Company, 2001.

[8] Christian Jacquemin. *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. PhD thesis, Université de Nante, Nantes, 1997.

[9] K. Kitamura and Y. Matsumoto. Automatic extraction of word sequence correspondences in parallel corpora. In *Proceeding of WVLC'96*, pages 79–87, 1996.

[10] Yoëlle S. Maarek, Daniel M. Berry, and Gail E. Kaiser. An information retrieval approach for automatically constructing software librairies. In *IEEE Transactions on Software Engineering*, volume 17(8), pages 800–813, aug 1991.

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

[12] Sayori Shimohata, Toshiyuki Sugio, and Junji Nagata. Retrieving collocations by co-occurrences and word order constraints. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 476–481, Madrid, Spain, July 1997.