TRAITEMENT AUTOMATIQUE

DU LANGAGES NATUREL

Isabelle STÉVANT – Master2 MSB - 2010-2011

Table des matières

I. INTRODUCTION

Depuis l'avènement du séquençage automatique, et au moment même où de nouvelles méthodes de séquençage nouvelle génération (NGS) sont mises en place, la quantité de publications issues de résultats en génomique est telle qu'il devient de plus en plus difficile d'en faire la lecture. Des méthodes de traitement automatique des langues, c'est à dire d'extraction d'information, pourraient aider à traiter ce flux de données.

A . Traitement automatique du langage naturel

Le Traitement automatique du langage naturel (TALN) ou Traitement automatique des langues (TAL) est une science à la croisée de plusieurs disciplines : la linguistique, l'informatique et l'intelligence artificielle (ou informatique cognitive). Il s'agit de mettre en place des programmes informatiques capables de traiter tous (ou presque) les aspects du langage humain. Dans le domaine de la génomique et de la protéomique, les objectifs principaux seraient d'extraire des relations moléculaires telles que les interactions ADN/Protéines et Protéine/Protéine afin de créer des graphes d'interaction entre ces entités de manière automatique.

Le traitement automatique des langues se décline en plusieurs niveaux de difficulté. Du niveau le plus faible au niveau le plus élevé, nous retrouvons l'approche par morphologie (forme des mots), l'approche syntaxique (combinaison des mots en énoncées), la sémantique (sens des mots/des énoncés), et enfin l'étude pragmatique (interprétation selon le contexte). Le premier niveau, la morphologie, n'en ai pas moins difficile, car pour comprendre la morphologie d'un mot, il faut tout d'abord savoir ce qu'est un mot. De nos jour, la définition de ce qu'est un mot ressort plus du ressentit que d'une définition propre. Pourtant tout le monde s'accorde sur le fait que la définition d'un mot est évidente.

On peut distinguer plusieurs types de mots :

- Le mot « graphique » : il s'inscrit entre deux espaces blanc. Ce découpage est conventionnel et est inscrit dans notre culture.
- Le mot « sémantique » : il s'identifie par sa signification. Il peut s'agir d'un bloc de signification (monème) ou d'unité plus petites (sème).
- Le mot « lexicale » : c'est autrement dit le mot figurant dans un dictionnaire de la

langue. Il diffère du mot graphique car il est ramené à sa forme initiale (singulier pour les noms, infinitif pour les verbes...). On l'appelle aussi lemme.

L'approche désormais classique de Ferdinand de Saussure (1916) associe à un mot deux dimensions : le signifiant (son) et le signifié (sens). À un même signifiant on peut associer plusieurs signifiés. Ceci révèle de nouveaux problèmes : celui de la polysémie et de la synonymie, mais encore la composition, c'est à dire l'association de plusieurs mots de sens différents pour créer un nouveau sens (exemple : porte-feuille).

Passons à la notion de syntaxe. Il s'agit d'étiqueter les mots en fonction de leur catégorie (nom, verbe, adjectif...). Cette approche permet de percevoir des relations entre les mots grâce à un verbe qui traduit une action. Cela dit, cette approche mène à de nombreuses ambiguïtés causées par exemple par des mot polyfonctionnels (exemples : porte, voile, brise...). Plusieurs outils ont été mis en place afin de déterminer au mieux la syntaxe d'un texte. On trouve des étiqueteurs en *Parties-du-discours* ou PoS (*Part-of-Speech*), déclinés pour plusieurs langues et qui désambiguïsent avec un taux de réussite supérieur à 95%, et des étiqueteurs syntaxiques mais peu précis et adaptés à peu de langages.

L'approche sémantique essaie de déterminer les relations en s'appuyant sur le sens des mots tout en prenant en compte la complexité du langage (synonymie, l'antonymie, l'hyperonymie...). Plusieurs théories ont été proposées mais elle s'avèrent peu compatibles entre elles. L'approche sémantique est surtout utile pour constituer des ontologies, qui permettent d'ordonner un ensemble connaissances grâce à un réseau de relation entre les entités.

B . Intérêt des couples nom-verbes

Il est clair que la relation entre un nom et une action est très riche en information et permet de d'enrichir le sens du nom. Prenons un exemple qui se rapporte au domaine de la génomique. Si l'on extrait d'un corpus constitué de publications portant sur une protéine dont on souhaite connaître la fonction, et que l'on extrait en majorité le couple nom-verbe suivant : protéine – inhibe, on peu sans trop se tromper en déduire que cette protéine a un effet négatif dans la voie métabolique où elle intervient.

II. MATÉRIEL ET MÉTHODE

Ainsi, en tentant d'extraire des relations nom-verbe dans un corpus de texte, nous allons tenter de mettre en avant le contenu informatif de ce corpus.

A . Constitution du corpus de texte

Le corpus est constitué de résumés d'articles scientifiques extraits de la base de donnée PubMeb. En théorie, chaque résumé fait environ 250 mots et est écrit en langue anglaise. Afin d'extraire des relations nom-verbes cohérentes, les résumés doivent traiter du même sujet. Pour ce faire, nous avons choisi comme mots clés : « caffeine, health, effects, men ». Nous avons extraits de cette requête 570 résumés.

Le corpus, après avoir été nettoyé de toutes les parties indésirables contenues dans le fichier de sortie de PubMed, est étiqueté en utilisant le logiciel TreeTagger, qui utilise les PoS et les lemmes. Voici un extrait de la sortie de TreeTagger :

MOT	TAG	LEMME
Overall	RB	overall
,	,	,
the	DT	the
results	NNS	result
from	IN	from
the	DT	the
three	CD	three
analyses	NNS	analysis
show	VBP	show
that	IN	that
caffeine	NN	caffeine
consumption	NN	consumption
may	MD	may
have	VB	have
benefits	NNS	benefit
for	IN	for
performance	NN	performance
and	CC	and
safety	NN	safety
at	IN	at
work	NN	work
.	SENT	.

B . Approche d'extraction des couples nom-verbe

L'approche choisie pour cette étude est très naïve. Elle consiste à extraire un couple nom-verbe si un verbe est présent dans une fenêtre de 3 mots suivant le nom. Dans cette approche, nous ne prenons pas en compte les constructions de phrases tels que les formes passives.

Dans un premier temps, les noms, c'est à dire tous les mots étiquetés avec « NN, NNS et NP », sont extrait du corpus et stockés dans un tableau associatif, où chaque clé correspond au numéro de la ligne où se trouve le nom. Quand le lemme du nom est connu, on le stock dans le tableau, mais s'il est « unknown », c'est le mot qui est stocké.

Dans un second temps, on extrait de la même manière tous les verbes, soit tous les mots taggués par « VB*. », * représentant une lettre ou rien. Comme précédemment, si le lemme est connu, on le stocke, s'il ne l'est pas, c'est le mot que l'on stocke. On gardera en plus des lemmes, les tag des verbes. Afin de ne garder que l'information essentielle, on va regarder de quelle manière sont conjugués les verbes. S'ils sont conjugués avec un auxiliaire (have ou be), on enlèvera cet auxiliaire pour ne garder que le verbe.

Enfin, on prend chaque mot de notre tableau et on regarde dans une fenêtre de 3 mots après lui si on trouve un verbe. Si oui, on stocke le couple dans un tableau.

C . Outils d'extraction

Les outils développés à cet occasions sont composé d'un script shell (*launch_corpus_analysis.sh*) qui prend en entré un fichier brut issu de PubMed. Il enlève toutes les parties inutiles et lance l'étiquetage par TreeTagger. Une fois le corpus étiqueté, il lance le script perl d'extraction des couples nom-verbes (*get_relation_NounVerb.prl*). Enfin, il affiche un résulé de l'analyse.

Voici un exemple de la sortie du script *launch_corpus_analysis.sh :*

```
zazooo@machin-gris:~$./launch_corpus_analysis.sh pubmed_result_caffeine_health_effect_men.txt
Ce corpus est composé de 570 textes.

Phase d'étiquetage du texte
        reading parameters ...
        tagging ...
151000   finished.

Ce corpus se compose de 129362 mots, dont 43629 noms et 15758 verbes, soit 33.73% de noms communs,
12.18% de verbes et 8.14% mots non reconnus.
```

```
Top 10 des couples nom-verbe:

87 | study - be
53 | result - suggest
51 | study - examine
42 | caffeine - increase
38 | result - indicate
37 | study - have
29 | consumption - associate
25 | finding - suggest
25 | result - show
24 | caffeine - associate

8955 couples nom-verbes uniques ont été détectés, soit 12009 au total.
```

III. RÉSULTATS

Comme le montre la sortie du script montrée un peu plus haut, nous avons identifié 8955 couples nom-verbes uniques, et 12009 couple au total. Le corpus étant composé de 15758 verbes, on peut en dire que 76% des verbes sont associés à un nom.

87	study	be
53	result	suggest
51	study	examine
42	caffeine	increase
38	result	indicate
37	study	have
29	consumption	associate
25	finding	suggest
25	result	show
24	caffeine	associate
24	study	investigate
22	caffeine	have
21	study	determine
21	study	suggest
20	datum	suggest
19	caffeine	improve
18	caffeine	be
18	caffeine	do
17	caffeine	produce
17	study	show

16	coffee	associate
15	study	conduct
14	caffeine	use
14	supplement	contain
12	caffeine	appear
12	caffeine	consume
12	finding	indicate
12	study	assess
12	study	use
11	level	increase
11	microM	increase
11	sample	collect
11	study	demonstrate
11	study	indicate
11	study	test
10	analysis	show
10	caffeine	compare
10	caffeine	reduce
10	coffee	have
10	objective	examine

10	rat	train
10	study	evaluate
9	caffeine	result
9	caffeine	show
9	coffee	increase
9	datum	collect
9	objective	investigate
9	result	support
9	subject	ingest
9	woman	age
8	Ca(2+	spark
8	caffeine	affect
8	caffeine	decrease
8	caffeine	enhance
8	caffeine	influence
8	change	observe
8	consumption	assess
8	consumption	have
8	effect	appear
8	evidence	suggest

Ces tableaux correspondent aux 60 premiers résultats. La première colonne correspond au nombre d'occurrence de chaque couple, la seconde colonne correspond aux noms, et enfin la troisième colonne aux verbes. En jaune, on peut voir le contenu le plus informatif.

On remarque que les couples nom-verbes les plus représentés sont en majorité des couples peu informatifs, mais compte-tenu de l'origine du corpus (résumé d'articles scientifiques) il est tout à fait normal de les trouver en grand nombre. On remarque d'autre part la forte proportion de couples contenant les mots « caffeine » et « coffee », ceci est d'autant plus normale qu'il s'agit du thème principal de notre corpus.

IV. Discussion

Au delà de ces résultats somme toute assez cohérents, cette approche comporte beaucoup de biais qui peuvent être améliorés de plusieurs manières. Commençons déjà par évaluer la qualité de l'étiquetage.

A . Qualité de l'étiquetage du corpus

Selon Helmut SCHMID, père de TreeTagger, le taux de réussite de l'étiqueteur est évalué à 94% soit 6% d'erreur d'étiquetage. Si on se base sur ces chiffres, il y aurait à peu près 7760 mots mals étiquetés dans notre corpus.

B . Limites de l'approche

Outre les erreurs d'étiquetage, avec notre approche qui consiste à extraire les noms et les verbes de leur contexte, nous ne prenons pas en compte la structure même des phrases dans lesquelles se trouvent nos couples. En effet, il est fort probable que l'on aie formé des couples constitué du dernier mot d'une phrase et du premier verbe de la suivante, comme dans cet exemple :

```
...documented in the litterature. It is associated with...
```

Ici, le couple qui sera formé est [litterature – associate]. En effet, l'auxiliaire « is » est enlevé par notre analyse, et si l'on regarde 3 positions après le nom « litterature », on trouve un verbe « associate ». Il faudrait donc prendre en considération de découpage du corpus en phrases et extraire les couples pour chaque phrase (groupe de mot compris entre deux « . »)

D'autre part, nous ne prenons pas en compte dans notre analyse les phrases passives, où le nom associé au verbe se trouve après ce dernier. Ceci va augmenter les faux couples comme vu ci-dessus. Pour prendre en compte cette forme particulière de phrase, il faudrait chercher les mots qui traduisent ce genre de structure, tels que « by » par exemple.

On observe aussi la construction de couples constitués d'une lettre pour le nom, mais cela relève plus de l'étiquetage que de l'extracteur. Il faudrait rajouter une condition qui va chercher le nom précédent si le nom courant n'est composé que de 2 lettre maximum.

Enfin, une autre source d'erreur et non des moindres, lorsque nous trouvons deux noms très proches suivi d'un verbe, on va généré deux couples distincts alors que le premier nom n'a sans doute aucun rapport avec le verbe, comme ci-dessous :

```
...Caffeine dosage was based...
```

Deux couples vont être générés : [caffeine – base] et [dosage – base].

C . Regroupement des noms par familles

Si nous devions grouper les noms par famille, on s'intéresserait aux verbes auxquels ils s'associent. En effet, le verbe reflète la propriété du nom, donc si un groupe de nom partage le même verbe c'est qu'ils partagent cette même propriété.

Voici un test avec le verbe « inhibit » :

caffeine	depolarization
microM	dose-dependently
adenosine	heart
agent	histamine
cell	increase
dose	ingestion
PABA	Iso
receptor	La3+
theophylline	mM
A2a	muscle
agonist	NaCl
alkaloid	opener
Ca2+	permeability
CGS-21680	result
coffee	RMP
component	RR
concentration	smoking
D(2)Rs	terminal
day	xanthine

Les mots en surbrillance ont a priori bien un lien avec l'action « inhibit », mais on ne peut pas conclure que tous ces éléments inhibent, ils peuvent être inhibés. Une information supplémentaire telle que la présence de phrase sous forme passive pourrait nous renseigné sur la vrai propriété de ces mots.

V. Conclusion

Malgré toutes les limites observées sur notre méthode, il en ressort que les couples nom-verbe extraits sont majoritairement cohérent. Cela dit, le contenu informatif relevée par ces couples reste faible. De plus, nous ne tenons pas compte des négations, ce qui biaise l'interprétation des relations.

Si l'on veut extraire un maximum d'information d'un corpus de texte, l'idéal est de procéder à une recherche par approche sémantique comme avec les outils Syntex et Upery qui analyse non seulement la syntaxe d'un corpus mais aussi les relations sémantiques entre le verbe et le nom.

Annexe 1 :

Script shell *launch_corpus_analysis.sh*