Talat

Rapport : Terminologie et stylométrie,
Extracteur d’unité terminologique polylexicale

Virgile CARTIER et Martial PASTOR

MASTER 2
Langue & Informatique

Table des matières

Introduction

L'idée qui présidait à l'élaboration de ce système était particulièrement simple : il s'agissait d'extraire automatiquement des termes liés au *TAL* à partir d'un corpus de plusieurs articles talistes issu des années 2007 à 2013.

Dés lors, deux problèmes de taille devinrent évident, la structure de notre corpus – pas de fichiers *XML* dont les balises pourraient être exploités pour faciliter l'extraction de paragraphes potentiellement pertinents – à laquelle s'adjoint la nature même de la tâche de notre système. De fait, l'extraction de termes soulève une interrogation délicate : qu'est-ce qu'un *terme* ? Ou plus précisément, comment distinguer un *terme* d'un *mot* ?

Il faudrait pour cela proposer une différenciation d'ordre linguistique pour systématiser l'opposition des mots issus de la langue commune à leur équivalent *mots-termes*, inscrits alors dans une métalangue (ici, le métalangage du *TAL*, à mi-chemin entre l'informatique et le linguistique). Ainsi, nous y reviendrons plus en détail, les premières versions de *talat* nourrissaient l'ambition d'extraire tout terme aussi bien monolexical que polylexical. Dés lors, nous avons dû nous rendre à la navrante évidence de la complexité de cette première tâche : des mots aux termes, la frontière est par trop ténue, alors comment affiner notre extraction ? Comment garantir que celle-ci se cantonne au strict *terme* ?

Cette interrogation nous aura malheureusement mené à une impasse : la distinction entre ces deux unités semblables étaient particulièrement complexe à poser – une observation que nous développerons plus longuement ci-après. En somme, nous avons remarqué à travers une annotation manuelle des termes d'un article de notre corpus que la délimitation d'un *terme* n'est pas une mince affaire – aussi il n'était pas évident pour deux individus (deux individus familiarisés à la terminologie talistico-linguistique qui plus est) d'annoter les mêmes termes et, car une complexité ne vient jamais seule, de le faire avec la même délimitation ; nous avons noté par exemple une confusion entre un *terme* et son apparente complémentation, qu'elle soit adjectivale ou par un complément du nom. À cela s'ajoute un autre facteur, plus obscur encore, l'aspect *technolectal* inhérent à ces articles tend à favoriser le terme au mot. En définitive, ce sont là autant de raison qui nous ont poussé à concentrer notre système sur l'extraction de termes polylexicaux, une tâche aussi réaliste à réaliser qu'intéressante à analyser. En effet, cette délimitation évoquée ci-dessus a cela d'intéressant qu'elle ne se cantonne pas qu'à deux ou trois unités, nous pourrions considérer des termes plus longs ; au reste, *talat* en repère et s'ils ne seraient pas nécessairement *stricto sensu* des termes, ils seront au moins de tendance éminemment terminologique.

En ce sens, nous allons dans ce rapport explorer notre méthodologie pas à pas, de notre premier *proof of concept* à notre dernière mouture de *talat*, et

à chaque étape nous analyserons les résultats obtenus. En ce sens, ce présent rapport racontera chronologiquement l'histoire de *talat* de ses balbutiements à sa version finale.

Premier *POC*, la méthode growthrate

La première méthode que nous avons réalisé avait pour but premier de nous familiariser avec ce corpus. Nous l'avons évoqué, celui-ci est composé d'article en *txt* et non en *xml*. En ce sens, il nous fallait définir ce que *talat* était supposé extraire à l'échelle de l'article entier – ainsi, nous avons dans un premier temps exploité le tokenizer de la librairie *nltk* afin de faire ressortir une liste des tokens pour chaque article sur lesquels nous avons calculé le *growthrate*.

Les résultats que nous avons obtenus étaient particulièrement décevants car ininterprétables : il y avait beaucoup de bruits – beaucoup de pré-traitements s'avéreront nécessaires pour nettoyer notre corpus – et les unités extraites apparemment pertinentes n'avaient pourtant rien d'extraordinaires : *ontologie*, *syntaxe* ou autre *règle* ; et sur ceux-ci la mesure de *growthrate* ne pouvait pas être indicatrice de quoi que ce soit de singulier. En définitive, la tokenization seule s'avère insuffisante, il n'en ressort que des termes plats dont les résultats mesurés sont proprement impénétrables.

Cela étant dit, cette première approche a eu le mérite de nous écarter d'une route qui nous semblait pourtant cohérente – nous devions extraire non pas des tokens mais des candidats potentiels d'unité terminologique – cela suppose donc de dépasser le traitement token à token pour passer à un traitement article à article.

Aussi, afin de nous faire une meilleure idée de ce que *talat* devrait être en mesure d'extraire, nous avons choisi de nous intéresser à ces *termes* si élusifs en les annotant manuellement.

Annotation manuelle mais complexe

Nous avons décidé d'annoter le premier article de notre corpus[1] par simple praticité. Quant à l'annotation, nous n'avons pas suivi les méthodologies d'usage – nous nous sommes attelé à l'annotation de termes sans essayer de définir ce qui pourrait, devrait être considéré comme tel. Nous avons donc enfreint l'un des principes clés de l'annotation de données : la référence à une documentation claire et unanime. En l'occurrence, bien que

1 *recital-2007-long-001*

nous ayons un profil quelque peu semblable, les termes que nous avons relevés étaient sensiblement disparates.

Martial	Virgile
ensemble d'axiomes logiques	axiomes logiques
règle grammaticale **forte**	règle grammaticale
ensemble de relations binaires	relations binaires
synsets	**synsets**
syntagme nominal	**syntagme nominal**

Tableau 1: exemples d'annotations du premier article de notre corpus

Comme nous pouvons le constater ci-dessus, certaines annotations semblent évidentes, notamment celles concernant les unités terminologiques du TAL ou de la linguistique les plus habituelles. En tout et pour tout, nous avons annoté 130 termes, l'accord annotateur que nous avons calculé est de 0.44, c'est-à-dire plutôt mauvais.

De plus, nous pouvons constater un facteur intéressant vis-à-vis des termes en langue anglaise. Ceux-ci sont régulièrement calqués tels quels du pendant technolectal anglais ; des termes NLP deviennent des termes talistes sans passer par une quelconque traduction. Aussi, si c'est là extrêmement rare en linguistique – il s'agira vraisemblablement d'un terme utilisé pour établir un *pont* technolectal d'une langue à l'autre – c'est en revanche courant concernant l'informatique. Le TAL à mi-chemin entre ces deux disciplines héritera de ces deux traditions, si certains termes se traduisent (*machine learning* → *apprentissage machine*) d'autres sont adaptés (*token* → *tokenization*, *tokenizer* ; pourquoi pas fétiche, traduction de ce mot en langue française ?), d'autres encore vont être utilisés tels quels, exactement comme en langue anglaise (*chunk*, *chunking*). Enfin, si ce technolecte taliste varie d'une langue à l'autre, les termes utilisés changent également de l'écrit à l'orale, néanmoins concernant la tâche de *talat*, cela n'a, ici, que peu d'importance.

Cela dit, nous pouvons dés lors constater deux potentiels soucis : la délimitation des unités polylexicales, comme soulignée dans notre tableau 1, mais aussi la distinction entre mot et terme monolexical.

La délimitation des unités polylexicales soulève un problème intéressant quant à la frontière d'un terme, est-ce que les éléments complémentaires d'un terme comme un adjectif ou un complément du nom forge un tout nouvel élément ? D'un point de vue sémantique, nous serions tentés de dire que c'est raisonnablement le cas, néanmoins cela ne nous paraît pas nécessairement évident car, comme souligné par F. Neveu, un terme permet de référer à un objet (abstrait, factuel, observable ou non) d'une discipline donnée mais ce qui va procéder de son acceptabilité en tant que terme *officiel* sera son utilisation dans la littérature de cette science. Ainsi, si le sens porté par un terme

composé sera distinct du terme *souche* (qui peut-être lui même polylexicale au reste) pourra t-il entrer dans une quelconque nomenclature si utilisé dans peu d'article ? D'autant que le TAL est une vaste discipline et si un terme est exploité dans un article quelconque avec une pleine et pertinente indépendance de sens, il n'est pas garanti de le retrouver ailleurs, non pas par inconvenance mais plus prosaïquement car l'objet auquel il réfère ne serait plus discuté.

Pour ce qui est des termes monolexicaux, le TAL échoit des problèmes de la linguistique qui puise dans la langue commune pour parler du langage – de nombreux mots courants auront un sens *courant* et un sens *métalinguistique* et il n'est pas évident de situer la différence, à plus forte raison dans un article scientifique empreint d'une teinte technolectale.

Ces considérations posées, cela nous donne une idée plus concise des termes que *talat* devrait *a priori* être en mesure de dégager – une multitude de possibles : aussi, l'intérêt ne sera peut-être pas de correspondre sobrement à la définition stricte d'un terme mais plutôt à extraire un objet conceptuel qui aurait le potentiel d'en devenir un. Pour ce faire, nous allons utiliser une librairie *python, Rake*, pour relever ces candidats.

Deuxième POC, utilisation du système Rake

Rake est un système automatique d'extraction de mots-clefs à partir de document individuel utilisable via *python*. Les auteurs de cette librairie sont partis du constat que « *Despite their utility for analysis, indexing, and retrieval, most documents do not have assigned keywords. Most existing approaches focus on the manual assignment of keywords by professional curators who may use a fixed taxonomy, or rely on the authors' judgment to provide a representative list* » [RAK010]. Globalement, il s'agit là précisément de ce que nous souhaitons réaliser grâce à *talat*.

L'extraction automatique de mots-clefs travaillait à l'origine à l'échelle d'un corpus, *Rake* tâche quant à lui de fonctionner à l'échelle du document « *we focus our interest on methods of keyword extraction that operate on individual document* » [RAK010] aux auteurs de souligner par ailleurs que l'on peut facilement passer du document au corpus « *by operating on asingle document, these methods inherently scale to vast collections and can be applied in many contexts* » [RAK010]. Dans notre cas, nous ferons aisément fonctionner le système de manière itérative, traitant chaque article indépendamment.

Cette librairie est particulièrement simple à utiliser et requiert quelques paramètres adaptables « *the input parameters for RAKE comprise a list of stop words (or stoplist), a set of phrase delimiters, and a set of word delimiters. RAKE uses stop words and phrase delimiters to partition the*

document text into candidate keywords, which are sequences of content words as they occur in the text » [RAK010]. En d'autres termes, il s'agit de retirer les mots sémantiquement vides « *RAKE is based on our observation that keywords frequently contain multiple words but rarely contain standard punctuation or stop words, such as the function words or other words with minimal lexical meaning* » [RAK010].

Néanmoins, il nous semble essentiel de souligner que ce système est développé pour l'anglais, le potentiel multilingue reposera alors sur *nltk* ; les *stopwords* susmentionnés devront être modifiés dans la mesure où ils incluent des conjonctions de coordinations (comme *de* par exemple) qui participent à la construction de complément du nom que nous souhaiterons détecter. Aussi, nous utiliserons les *stopwords* proposés par *nltk* fr desquels nous en retirerons quelques uns.

Aussi, pour éviter une trop grande sélection de candidats potentiels par *Rake*, nous avons choisi d'utiliser le *GLAFF* pour filtrer les mots qui ne nous intéressent pas :

- tout candidat ne contenant pas un substantif en première position est exclu
- tout candidat avec des conjonctions de coordination, avec des prépositions (à l'exception de certaines à fonction multiples comme le "de", "à"), avec des verbes conjugués et avec des articles

Aussi, il nous faut noter que de ces mots, certains seront semblables à certains verbes et peuvent ainsi occasionner un léger flou car étant réportier autant comme verbe que substantif dans le *glaff* par exemple.

Les résultats que nous obtenons sont particulièrement encourageants, *sémantique des termes, termes candidats, unités terminologique, syntaxe de dépendance etc*. Ceux-ci restent cependant quelque peu gâtés par du bruit et d'autres problèmes un peu plus complexes : les adverbes en fin de postion sont ambiguës dans biens des cas *termes fréquemment* vs *termes extrait automatiquement,* nous avons de ce fait certains candidats surperflus.

Troisième POC, extraction des mots clés explicitement déclarés

Comme nous l'avons déjà évoqué, notre corpus est composé de documents en texte brut – il n'y a aucune structuration des données explicite. Il est particulièrement difficile d'identifier et d'isoler les parties pertinentes du texte sans balisage *XML*, il nous est impossible d'extraire le contenu de la balise correspondant à l'introduction, aux mots-clefs, à l'introduction ou encore à la conclusion qui sont pourtant les parties les plus pertinentes pour une extraction de candidats terminologiques potentiels.

Pour contrevenir à ce soucis, nous devons trouver des procédés inédits. En regardant de plus près nos articles, nous avons constater que certains articles présentaient une liste de *keywords* en langue anglaise avant de les présenter en langue française. Ainsi, en utilisant la librairie *langdetect* nous avons été en mesure d'extraire une série de mots-clefs en repérant les séquences langue anglaise → langue française → langue anglaise.

Globalement, les résultats que nous avons obtenus grâce à cette technique n'étaient pas sensationnels, néanmoins ils étaient tout particulièrement pertinents – il n'y avait aucun bruit pour venir parasiter les candidats potentiels.

Aussi, si cette méthode n'est pas d'elle même suffisante pour notre système, elle s'adjoint particulièrement bien à *Rake*. En effet, nous avons pu remarquer que 73 % des termes relevés ainsi l'étaient également par notre première version de *talat* ne prenant pas en compte le filtrage plus avancé avec le glaff. Dans la seconde version 100 % des termes étaient détectés. En ce sens, l'utilisation conjointe de ces deux méthodes nous a permis d'évaluer nos deux versions de *talat*.

Résultats

Nous n'avons malheureusement pas pu calculer les mesures d'usage sur l'intégralité de notre corpus en tant qu'il n'a pas été totalement annoté – aussi, seul le premier article que nous avions annoté aura pu faire office de comparaison. Il nous faut tout de même rappeler qu'il n'est pas nécessairement évident de déterminer précisément ce qui peut être considéré comme terme, nous en avons évoqué les potentielles raisons ci-dessus. En ce sens, ces mesures n'ont qu'un intérêt succinct mais elles offrent tout de même une bonne indication de l'opérabilité de la première version de *talat*.

Nous obtenons ainsi sur notre premier article annoté 0.52 de précision et, à y regarder de plus près, nous y trouvons des termes qui nous avaient échappés et qui trouvent tout de même une pertinence certaine à être considérés comme des candidats terminologiques.

Martial et **Virgile**	**Talat**
Axiomes logiques	Proche sémantiquement
Maximum-entropy tokenizer	Connectivité
Modules réseaux communicants	Génératif
Open nlp	Arité maximale
Wordnet	Mots clés
Sac de mots	Chunké
Contexte	Algorithme ascendant génératif

Tableau 2: termes annotés par les auteurs et par Talat dans le premier article du corpus

Au risque de nous répéter, les délimitations des unités polylexicales offrent un grand potentiel à l'acception terminologique. En revanche, nous avons pu constater que les unités monolexicales extraites étaient par trop perturbées par des mots communs. Devant ce bruit, nous avons pris la décision de nous limiter à l'extraction polylexicale. Regardons maintenant ce que celle-ci nous donne à l'échelle du corpus.

Mesure de similarité	**278**
Extraction de relation	**170**
État de l'art	**161**
Modèle de langue	**144**
Traitement automatique des langues	**129**
Moteur de recherche	**116**
Machine translation	**58**
Arbre de dérivation	**55**
Traduction automatique statistique	**54**
Réécriture de graphe	**51**

Tableau 3: Les 10 termes les plus extraits par le système Talat sur notre corpus

Pour ce qui est des termes les plus repérés, il est tout à fait intéressant de remarquer que ceux-ci rentrent tout à fait dans ce qui pourrait être considéré comme étant de l'ordre terminologique – à n'en pas douter, le système *talat* offre des résultats convaincants.

Aussi, nous pouvons maintenant associé notre première méthode de *growthrate* qui, couplée à ce traitement, devrait permettre d'illustrer l'avènement de tendance conceptuelle en TAL.

Reconnaissance automatique	
2007 - 2008	**0.0**
2008 - 2009	**Infini - apparition**
2009 - 2010	**2.0**
2010 - 2011	**0.25**
2011 - 2012	**2.0**
2012 - 2013	**1.0**

Tableau 4: growthrate du terme "reconnaissance automatique" extrait et calculé par talat

Nous avons sélectionné un terme au hasard sur le rendu obtenu. Nous pouvons constater ici que ce terme apparaît dans notre corpus en 2008 - 2009, que son utilisation augmente en 2009 - 2010 pour diminuer en 2010 - 2011 puis ré-augmenter en 2011 - 2012 pour enfin se stabiliser en 2012 – 2013. Nous pouvons aisément récupérer ces valeurs pour générer des courbes illustratives.

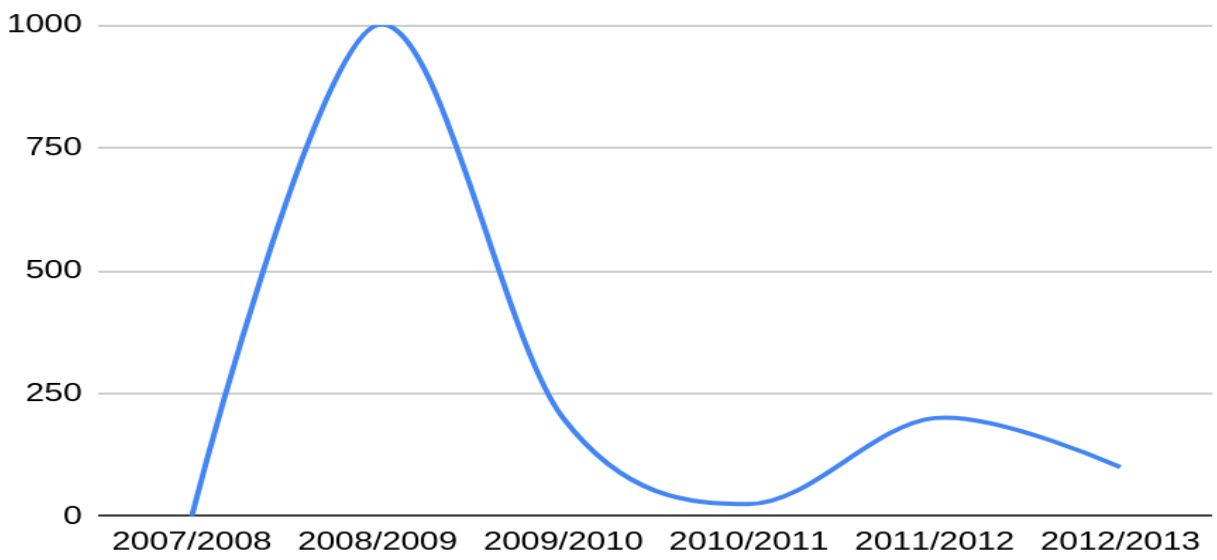

Figure 1: courbe du growthrate du terme « reconnaissance automatique »

Néanmoins pour pouvoir interpréter ces résultats de façon approprié, il nous faut également considérer la taille de nos articles dans nos corpus.

Année	taille en caractères
2007	1 798 925
2008	1 527 338
2009	**2 100 212**
2010	1 789 154
2011	**2 371 495**
2012	**2 084 719**
2013	1 685 546

Tableau 5: Taille des articles de notre corpus en caractères par année

Nous avons choisi de considérer la taille de nos articles en caractères et pouvons constater qu'en moyenne ceux-ci ne sont pas parfaitement équivalents. Néanmoins, leur différence n'est pas excessive et au regard de la courbe de growthrate du terme sélectionné, elle ne semble pas avoir joué de rôle décisif sur sa représentativité. Il n'est cela dit pas à exclure que cela pourrait avoir incidence sur d'autres termes extraits par *talat*.

Talat offre néanmoins une robuste extraction de *simili* termes qui, s'ils ne sont pas nécessairement considérables comme des unités terminologiques pures, peuvent du moins offrir une illustration concrète des concepts talistes discutés et explorés dans ce corpus. Illustration d'autant plus intéressante que celle-ci repose sur des données mesurés automatique par le système.

Au regard de ces résultats, il nous semble que *talat* rempli bel et bien ses prérogatives. Un obstacle néanmoins à cette observation : celles-ci ne sont pas totalement fixes – il se peut que certains termes extraits qui seraient à nos yeux considérables comme tels puissent sembler radicalement aberrants à d'autres talistes.

Perspectives et conlusion

Si l'étude d'année à année peut s'avérer intéressante, nous aurions pu concevoir *talat* de façon a ce que celui-ci intègre différentes échelles d'étude, mois à mois, trimestre à trimestre ou encore semestre à semestre, etc.

Notre abandon des termes monolexicaux peut être perçu comme un constat d'échec, nous n'avons pas réussi à trouver une solution convaincante pour catégoriser un candidat comme unité terminologique ou unité lexicale. Au reste, la tâche ne semble pas évidente.

Certains termes candidats incluant des adverbes peuvent générer du bruit dans nos résultats, il aurait fallu opérer un filtre cas par cas pour sélectionner qui de ceux-ci serait en mesure de composer une unité terminologique.

Enfin, nous aurions pu faire des analyses allant au-delà de l’approche embryonnaire de la fréquence d’apparition des termes. Nous aurions pu nous pencher sur des considérations sémantiques nous permettant de mesurer le poids d’un terme dans un article indépendamment de sa fréquence d’apparition.

Bibliographie

RAK010: Stuart Rose, Dave Engel, Nick Cramer, Wendy Cowley, Automatic keyword extraction from individual documents, 2010