S6Info AixMontperrin

Détection de néologismes
Rapport projet Langage Naturel

Gallina Ygor

Enseignants: Ramisch Carlos
Stafanini Marie-Héléne

1 Introduction

Identifier les mots nouveaux n'est pas une tâche aisée, cela commence par la définition de néologisme. Qu'est-ce qu'un néologisme ? Nous verrons qu'il n'y à pas de réponse claire à cette question.

L'objectif de ce travail consiste à extraire les néologismes d'un corpus, en utilisant les techniques du TAL.

Exemple d'entrée/sortie attendue :
`Un locavore et un islamophobe`
$\rightarrow \rightarrow \rightarrow \rightarrow$
`locavore`
`islamophobe`

Nous aborderons premièrement l'aspect linguistique de la recherche de néologisme, puis l'aspect informatique. Chacune de ces deux parties sera découpé d'abord par la recherche de non-néologismes, puis par la différenciation entre les coquilles et les "vrai" néologismes.

2 Linguistique

On utilisera "mot" pour représenter une unité de sens, et non une unité lexicale délimitée par des espaces.

2.1 Définitions

Les néologisme sont des mot qui entrent dans l'usage d'une langue, puis dans son lexique. Beaucoup de ces mots ne rentrent pas dans le lexique de la langue car ils sont un effet de mode et ne s'imposent pas dans le temps. La plupart des néologismes sont créés à partir de mot déjà existants.

Un néologisme est un mot nouveau ou apparu récemment dans une langue, le phénomène de création de nouveaux mots étant appelé, de manière générale, « néologie ». C'est un mot en cours de lexicalisation pourrait-on dire.

La linguiste T. Cabré propose plusieurs définitions sur ce qu'est un néologisme :

- Définition psychologique
  Un néologisme est un mot perçu comme nouveau par les usagers de la langue.

- Définition lexicographique
  Tout mot n'apparaissant pas dans un lexique d'exclusion est un néologisme. (donc un mot est un néologisme dans un contexte particulier, celui de la liste d'exclusion choisie)

- Définition diachronique
  Tout mot apparaissant dans un texte général et récent et qui ne faisait précédemment pas partie du langage est un néologisme. (encore ici par rapport à une définition du langage mais à un moment précis (introduction de temporalité))

Dans ces définitions on trouve toujours l'idée de néologisme par rapport à un lexique de cardinalité fini. Hors une langue vivante est par définition vivante, ce qui un implique un vocabulaire qui évolue et un vocabulaire possiblement infini. Définir un néologisme n'est pas si simple. La première définition, psychologique, me paraît la plus précise bien que la notion de "perception" ne puisse être définie formellement, ce qui fait qu'un néologisme est nouveau dans un référentiel qui n'est plus le lexique, mais l'humain, ce qui est plus vague. La définition diachronique introduit la temporalité comme nouveau référentiel en plus de celui du lexique. On pourrait formuler cette définition logiquement par

```
NON (faitPartieDuLexique(motCandidat, lexique))
        ET lexiqueCrééAvant(année, lexique)
```

de ce fait un mot est un néologisme par rapport à un lexique créé à un moment donné.

Il faut aussi définir la notion de nouveauté d'un mot. La nouveauté d'un mot est assez relative, un mot est nouveau dans une langue précise, à un moment précis. il n'y à pas de période au-delà de laquelle un mot n'est plus nouveau. Dés lors comment savoir quand un néologisme n'en est plus un ? C'est l'usage du mot qui le rend pérenne, quand les usagers de la langue ne perçoivent plus le mot comme nouveau, et que le mot n'est pas un effet de mode il sera lexicaliser. Si le mot n'est que peu utilisé, il ne sera lexicaliser, et retournera à l'état de non-mot.

Après avoir tenté de définir un néologisme, les linguistes distinguent différentes catégories de néologismes :

- Morphologique
  Le néologisme de forme est un nouveau mot, qui apparaît dans le lexique d'une langue, par transformation d'un mot déjà connu ou par création d'un lemme.

- Syntaxique
  Utilisation d'un mot connu, dans une partie du discours autre que celles connues. Par exemple utiliser un mot qui est seulement un nom, en tant qu'adjectif, ou verbe ...

- Néosémique
  Utilisation d'un mot connu, dans une partie du discours qui existe déjà, dans un sens nouveau. Par exemple « virus » qui fût un temps était utilisé seulement dans le domaine médical, est maintenant employé en informatique et ont des sens différents.

2.2 Différencier les candidats

Après avoir écrémé les mots connus du corpus d'entrée grâce à une liste d'exclusion, il faut maintenant décider si un mot est vraiment un néologisme ou pas.

Un non-néologisme peut être un mot inconnu de la liste d'exclusion, une coquille (un mot mal orthographié), un noms propre, un mot d'une langue étrangère, ...

Une flexion d'un mot connu peut être reconnu en ayant un dictionnaire fléchie, ou en lemmatisant le mot.

Une coquille peut être différencié avec un correcteur orthographique. Un néologisme pourrait être très proche d'un mot connu et serait donc catégorisé comme non-néologisme par un correcteur orthographique, c'est un des cas limite qu'un humain aurait peu de mal a traiter, mais qui est difficile à automatiser (cf. Définition psychologique).

Un mot d'une **autre langue** sera toujours considéré comme un néologisme, a moins de de le traiter comme une coquille et utiliser un correcteur orthographique pour plusieurs langues. Ce qui est un peu lourd. On pourrait imaginer de ne faire cela qu'avec l'anglais car les mots étrangers introduits sont pour la plupart anglais.

Il peut aussi y avoir des mots qui appartiennent au langage mais ne sont pas dans un dictionnaire car ce sont des mots composés par exemple en allemand le mot "verjaardagstaart" qui signifie gâteau d'anniversaire est composé de "verjaadag" -anniversaire- et "tart" -gâteau-, qui sont tout deux dans le dictionnaire, mais "verjaardagstart" n'y est pas car son sens est induit.

3 Informatique

La seule information que l'on ai sur les néologisme est qu'ils ne sont pas dans le dictionnaire. Il faut donc trouver tout ce qui n'est pas un néologisme.

3.1 Outils existants

Plusieurs outils de détection sont mentionnée dans la littérature, nombre d'entre eux sont introuvable. Des analyseurs de partie du discours peuvent aussi jouer ce rôle (cf. Étiquetage morpho-syntaxique pour des mots nouveaux, Falk et al.). Mais il existe des base de donnée (inaccessible au grand public) qui aide l'identification.

Base de données :

MorDebe	Base de donnée de 135.000 lemmes et 1.5M formes fléchies (portugais)
NeoTrack	Requête sur la base de donnée MorDebe
Neologia	Base de donnée de néologisme (français)
Morfetik	Base de donnée de 1M formes fléchies (français)

3.2 Prétraitements

Avant tout traitement, il faut pré-traiter le corpus pour avoir un format simple à utiliser. Ici on tokenise le corpus d'entrée avec `tokeniser.perl` de Josh Schroeder, puisque l'on veut traiter chaque mot. On applique une expression régulière pour avoir un mot par ligne, pour n'avoir qu'a traiter le caractère de séparation `'\n'`. On met les tokens en minuscule grâce à `lowercase.py` du `mwetoolkit`.

3.3 Élimination des mots connus

La suppression de "stopwords" permet d'avoir à traiter moins de mots dont on connaît le sens et qui a priori ne sont pas des néologismes (la, le, les, etc...).

Il y a eu plusieurs itérations de ce traitement dans l'application :

- Utilisation de la commande `wordnet`
  Le dictionnaire wordnet contient un nombre conséquent de mots. L'inconvénient est que cela ne fonctionne que pour l'anglais.

- Utilisation d'un dictionnaire
  Un **dictionnaire traditionnel** est une bonne liste d'exclusion et est modulaire, mais les formes fléchies des mots n'en font pas partie
  Un **dictionnaire des formes fléchie** est une meilleure base. Le dictionnaire DELA paraît une bonne option, il contient 700.000 entrées, et est donc assez complet. A titre de comparaison le LEFFF contient "seulement" 400.000 entrés.

- L'utilisation d'un arbre des préfixes, serait une bonne optimisation pour la vitesse d'exécution par rapport à la commande `grep`. La recherche d'un mot dans le dictionnaire se fait en $o(wordLength)$ dans un arbre des préfixes, tandis que qu'elle s'effectue en $o(tailleDict)$ avec la commande `grep`.

3.4 Différenciation des candidats

3.4.1 Coquilles

Pour faire la différence entre une coquille et un néologisme, la solution la plus appropriée semble d'utiliser un correcteur orthographique. Peter Norvig à écrit une très bonne introduction à la correction automatique (http://norvig.com/spell-correct.html), dont le code est open-source et simple.

Ce programme utilise les éditions de la distance d'édition (ajout, suppresion, transposition, remplacement) sur un mot pour essayer de le faire correspondre à un mot connu, si il n'y à aucun candidats (mots connus) on réitère l'opération.

Le problème est que le programme corrige le mot même si il est assez éloigné du mot d'origine (possiblement faux), par exemple le mot 'swag' est corrigé en 'sa', alors que 'swag' est un néologisme d'emprunt à la langue anglaise et devrait donc ne pas être corrigé. Dans cet exemple le 'w' et le 'g' ont été supprimés. Ors supprimer un 'w' et un 'g' n'est pas anodin, ce sont des lettres qui apparaissent assez peu. Ajouter un 'w' dans un mot est une faute moins courante que doubler une consonne ou ajouter un 'e' ou un 's' à la fin d'un mot.

Une première piste pour pallier à ce problème serait de ne pas tester toute les éditions possibles pour chaque mot mais de tester certaines erreurs courantes en utilisant le fuzzy pattern matching (correspondance de motif approximative) avec la commande `agrep`.

Comment modifier un mot(potentiellement mal orthographié) pour lui ajouter des expressions régulières ?
Exemple avec le mot mal orthographié : 'anticonstitutionelemen'
On pourrait tenter de :
Doubler les consonnes d, h, l, t, m, n, s, r, p, f, c
Remplacer les 'a' par des 'e' et inversement
Ajouter des 's' et 'e' terminaux
...
Pour obtenir une expression régulière qui pourrais ressembler à ça :

```
'[ae]ntt?iconstt?itt?utt?ionn?[ea]ll?[ea]me[na]t?e?s?'
```

Faire des expressions régulières pour les fautes les plus rencontrés serait contraignant et peu modulaire.

Une autre approche serait de donner un poids à chaque édition en fonction du type d'édition et de la lettre modifiée. Ensuite utiliser le poids pour décider si le mot est corrigeable ou non, auquel cas c'est un néologisme.
Mais comment modéliser cela dans un langage de programmation?
Faire un cas pour chaque faute serait laborieux et peu lisible.

Une approche probabiliste semble être une bonne approche.

Une édition change un caractère dans le mot d'origine. On pourrait regarder si le n-gramme de caractères qui englobe le caractère édité, apparaît souvent dans un corpus.

Par exemple le mot mal orthographié 'bonjor', et l'édition 'bonjour', qui a ajouté un 'u', on prend le 2-gramme 'ou' qui apparaît souvent dans la langue française aura plus de chance d'être une édition "juste" que 'bonjowr'. Bien sûr, un seul 2-gramme n'est pas suffisant pour évaluer de la justesse d'une édition. Si l'on fait la moyenne des poids des 2,3,4-grammes le résultat de la pondération sera plus juste. Choix des 2,3,4-grammes :

b	o	n	j	o	w	r
				o	w	
					w	r
				o	w	r
			j	o	w	r

Le corpus d'entraînement ne doit pas être trop récent, pour ne pas contenir des néologismes pas encore lexicalisé, ni trop éloigné de l'époque actuelle pour qu'il représente la langue utilisé par le corpus dont on veut extraire les néologismes. Compter le nombre d'occurrence des n-grammes des mots est trivial, il faut extraire la sous-chaîne de taille n commençant à chaque caractère du mot, si elle est définie. Le moins trivial est de trouver la formule pour calculer le poids de l'édition.
Sachant que l'on ne peut comparer des corpus de taille différente, il faut trouver un moyen de faire correspondre les nombre d'occurrence entre les 2-grammes, 3-grammes, ... Une solution pourrait être de ramener chaque valeur entre 0 et 1 et d'en faire la moyenne, puis de pondérer cette moyenne en fonction du caractère édité. Par exemple en suivant l'idée de la rareté d'une lettre comme dans le Scrabble. Éditer une lettre chère devrait être un événement peu probable.

Ainsi, on peut restreindre le nombre de candidats à la correction d'un mot, et éventuellement le rendre nul.

3.4.2 Nom propre

L'identification des nom propres est un problème résolu du TAL car des programmes avec une fiabilité $> 90\%$ ont été dévellopé. Shlomi Babluki en parle dans son blog (https://thetokenizer.com/2013/08/25/how-to-easily-recognize-peoples-names-in-a-text/). Identifier et retirer les noms propre en même temps que les stop-words permetrait d'éviter d'avoir à les traiter avec les coquilles et de générer de faux positifs.

3.5 Création du corpus de test

Le corpus de test doit être assez contemporain pour contenir des néologismes. Les texte contemporain sont en général des textes journalistiques. Les journaux susceptible de contenir des néologismes sont les journaux dématérialisés sur internet. Par exemple les magazine vice.com, konbini.com, liberation.fr, lemonde.fr.

Conclusion

Identifier les néologismes d'un texte n'est pas trivial. Aucune définition claire n'existe à ce jour, ce qui rend l'automatisation complexe. La méthode de différenciation des coquilles n'est pas parfaite, il faudrait l'améliorer ou partir sur une autre base. D'autres méthodes sont abordées dans la littérature. Pour simplifier le processus, il faudrait posséder un lexique complet de la langue, et un texte en entrée qui ne contienne aucune faute d'orthographe, de cette manière avec la seule définition "Tout mot n'apparaissant pas dans un lexique d'exclusion est un néologisme", l'identification de néologisme serait plus simple.

Bibliographie

S. Mejri : "Néologie et unité lexicale : renouvellement théorique, polylexicalité et emploi"

P. Drouin , A. Paquin , N. Ménard : "Extraction semi-automatique des néologismes dans la terminologie du terrorisme"

B. Sagot, D. Nouvel, V. Mouilleron, M. Baranes : "Extension dynamique de lexiques morphologiques pour le français à partir d'un flux textuel"

N. Lemaire, M. Van Campenhoudt, 2008 : "Détection et classification de néologismes : une expérience didactique"

M. Janssen : "Orthographic Neologisms"

I. Falk, D. Bernhard, C. Gérard, R. Potier-Ferry, 2014 : "Étiquetage morpho-syntaxique pour des mots nouveaux"

E. Cartier, J.-F. Sablayrolles : "Nouvelles technologies, nouveaux modèles linguistiques et néologie"

I. Falk, D. Bernhard, C. Gérard, 2014 : "From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers"

M. Janssen : "NeoTag: a POS Tagger for Grammatical Neologism Detection"

F. Issac : "Cybernéologisme : Quelques outils informatiques pour l'identification et le traitement des néologismes sur le web"