

Construction de lexiques pour l'extraction des mentions de maladies dans les forums de santé

Elise Bigeard^{1 2}

¹Univ. Lille, CNRS, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France

²Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, team ERIAS, UMR 1219, F-33000 Bordeaux, France

29 juin 2017

Objectif

Ressources morphologiques

Ressources sémantiques

Méthodes distributionnelles

Meta

Objectif

Extraire et identifier les maladies dans les posts de forums santé

Objectif

j'aimerais savoir si c'est possible de prendre de temps en temps un xanax 0,25mg après avoir allaité son bb, car j'ai une **phobie sociale**

phobie sociale – agoraphobie

Corpus

- ▶ 60 000 posts
- ▶ post typique 50 à 150 mots
- ▶ plutôt bien orthographié mais avec des abréviations et fautes

Expressions recherchées

- ▶ je viens de commencer un traitement pour **anxiete generalisee**
- ▶ j'ai eu une crise de **spasmo** des plus violentes
- ▶ J'ai d'abord été diagnostiqué **dépressif**
- ▶ Depuis 4 ans je suis **accro** au tramadol
- ▶ je ne **dormais plus** la nuit
- ▶ j'ai en fait peur que le zyprexa m'ai détruit de l'intérieur, je ne me reconnais plus, plus rien ne m'intéresse, je n'arrive plus à réfléchir, ni à imaginer..

Cas d'usage

troubles de l'humeur

6 000 posts dans la catégorie "anti-dépresseurs et anxiolitiques"
2ème plus grosse catégorie

On recherche 23 maladies associées aux anti-dépresseurs dans 100
posts annotés manuellement

Travaux Connexes

- ▶ Consumer Health Vocabulary (anglais) [Zeng and Tse, 2006]
- ▶ Paraphrases pour les termes composés (français) [Grabar and Hamon, 2014]
- ▶ Lexique sur le cancer du sein (français) [Eholié et al., 2016, Messai et al., 2006]

Noms canoniques de maladies

uniquement les 23 maladies dans leur forme canonique

baseline & point de départ ("seed")

agoraphobie, boulimie, schizophrénie, anxiété généralisée ...

Noms canoniques de maladies

termes canoniques uniquement

évaluation sur 100 messages annotés manuellement

lexique	TP	précision	rappel	f
<i>seeds</i>	98	0.859	0.388	0.535

faux positifs : maladies qui sont aussi un mot du langage courant
"je panique", "peur panique"

Objectif

Ressources morphologiques

Ressources sémantiques

Méthodes distributionnelles

Meta

Lexique.org

termes de la même famille morphologique

anxiété : anxieux, anxiolytique, anxieusement

nervosité : nerveux, nervure

lexique	TP	précision	rappel	f-mesure
<i>seeds</i>	98	0.859	0.388	0.535
<i>lexique.org</i>	120	0.869	0.476	0.615

Objectif

Ressources morphologiques

Ressources sémantiques

Méthodes distributionnelles

Meta

Wikidata

base sémantique de Wikipédia

24+ millions d'objets

8 millions d'objets dans le domaine médical

7 700 noms de maladies extraits

Wikidata



WIKIPÉDIA
L'encyclopédie libre

Accueil
Portails thématiques
Article au hasard
Contact

Contribuer

Débuter sur
Wikipédia

Article [Discussion](#)

Agoraphobie

(Redirigé depuis [Peur sociale](#))



Des informations de cet article ou section devraient être

Améliorez sa **vérifiabilité** en les **associant par des références** à l'aide d'a



Ne doit pas être confondu avec [Ochlophobie](#).

L'**agoraphobie** (du **grec ancien** *ἀγορά* / *agorá* (« place publique », « assemblée ») et *φόβος* / *phóbos*) manifeste par un sentiment d'insécurité dans les lieux publics ou les vastes espaces et par la peur

Nombre de gens qui se disent agoraphobes sont en fait **démophobes** : ce ne sont pas les lieux ouve

Wikidata



Item [Discussion](#)

agoraphobia (Q174589)

A phobic disorder involving the specific anxiety about being in a place or situation where escape is difficult or embarrassing [edit](#)
 where help may be unavailable.

fear of open spaces (finding)

[In more languages](#) [Configure](#)

Language	Label	Description	Also known as
English	agoraphobia	A phobic disorder involving the specific anxiety about being in a place or situation where escape is difficult or embarrassing or where help may be unavailable.	fear of open spaces (finding)
French	agoraphobie	No description defined	agoraphobe peur sociale
Spanish	agorafobia	No description defined	miedo a los espacios abiertos fobia a los espacios abiertos
German	Agoraphobie	Krankheit	

- [Main page](#)
- [Community portal](#)
- [Project chat](#)
- [Create a new item](#)
- [Item by title](#)
- [Recent changes](#)
- [Random item](#)
- [Query Service](#)
- [Nearby](#)
- [Help](#)
- [Donate](#)

- [Tools](#)
- [What links here](#)
- [Related changes](#)
- [Special pages](#)

Wikidata

termes associés au même code de maladie qu'une seed

dépendance : toxicomanie, polyconsommation

schizophrénie : schizophrène, schizofrene, catatonique

bipolaire : bipolarité, psychose maniaco-dépressive, folie circulaire

insomnie : parasomnie, somnanbulisme, terreur nocturne,

cauchemar pathologique

Wikidata

lexique	TP	précision	rappel	f-mesure
<i>seeds</i>	98	0.859	0.388	0.535
<i>lexique.org</i>	120	0.869	0.476	0.615
<i>wikidata</i>	100	0.862	0.396	0.543

Jeux de Mots¹

réseau lexical créé à partir d'un jeu en ligne

relations entre les mots possèdent un type et un poids

¹[Lafourcade, 2007]

Jeux de Mots

termes de la même famille morphologique

dépression : déprimé, anti-dépresseur, dépressuriser

insomnie : insomniacque, somnifère, sommeil, hypersomnie,
somnanbulisme

Jeux de Mots

lexique	TP	précision	rappel	f-mesure
<i>seeds</i>	98	0.859	0.388	0.535
<i>lexique.org</i>	120	0.869	0.476	0.615
<i>wikidata</i>	100	0.862	0.396	0.543
<i>jdm morpho</i>	119	0.856	0.472	0.608

Jeux de Mots

30 premiers termes associés à la seed, toutes catégories confondues

phobie : crainte, terreur, dégoût, aversion, panique, araignée

suicide : se donner la mort, s'ouvrir les veines, overdose,
médicament, anti-dépresseurs, rater

boulimie : hyperphagie, frénésie, appétit, avidité, maladie

Jeux de Mots

lexique	TP	précision	rappel	f-mesure
<i>seeds</i>	98	0.859	0.388	0.535
<i>lexique.org</i>	120	0.869	0.476	0.615
<i>wikidata</i>	100	0.862	0.396	0.543
<i>jdm morpho</i>	119	0.856	0.472	0.608
<i>jdm</i>	132	0.469	0.523	0.495

Objectif

Ressources morphologiques

Ressources sémantiques

Méthodes distributionnelles

Meta

Méthodes distributionnelles

Brown clustering² & Word2Vec³

explorent le corpus pour faire émerger des termes proches

basés sur les relations paradigmatiques entre les mots

²[Brown et al., 1992, Liang, 2005]

³[Mikolov et al., 2013a, Mikolov et al., 2013b]

Brown clustering

500 clusters

boulimie, insomnie, agoraphobie :
insomnie, déprime, agoraphobie, migraine, cauchemar, tristesse,
sommolence, vomissement, tachycardie, tag, boulimie,
spasmophilie, pleurs, déréalisation...

dépendance :
deuil, **sevrage**, **addiction**, voyage, **rémission**, descente, ame,
substitution, évaluation, métier, caractère, croissance...

Brown clustering

30 premiers termes du cluster auquel la seed appartient

lexique	TP	précision	rappel	f-mesure
<i>seeds</i>	98	0.859	0.388	0.535
<i>lexique.org</i>	120	0.869	0.476	0.615
<i>wikidata</i>	100	0.862	0.396	0.543
<i>jdm morpho</i>	119	0.856	0.472	0.608
<i>jdm</i>	132	0.469	0.523	0.495
<i>brown</i>	105	0.541	0.416	0.470

Word2Vec

cbow, bigrammes, +-10 mots

crise d'angoisse :

attaque, spasmophilie, agoraphobie, panique, spasmo, anxiété généralisée, tétanie, impulsion, tachycardie, anxiété, larme, dépersonnalisation, déprime, somatisation, déréalisation ...

bipolaire :

borderline, psychotique, tag, trouble bipolaire, thymoregulateur, phobie sociale, schizo, dépressif, hypocondrie, trouble anxieux, schizophrène, réactionnel, anxiété généralisée, agoraphobie, bipolarité, anxios ...

Word2Vec

30 premiers termes associés à une seed

lexique	TP	précision	rappel	f-mesure
<i>seeds</i>	98	0.859	0.388	0.535
<i>pos</i>	120	0.869	0.476	0.615
<i>wikidata</i>	100	0.862	0.396	0.543
<i>jdm morpho</i>	119	0.856	0.472	0.608
<i>jdm</i>	132	0.469	0.523	0.495
<i>brown</i>	105	0.541	0.416	0.470
<i>w2v seeds</i>	109	0.480	0.432	0.455

Word2Vec

30 premiers termes associés à une seed + un mot mal orthographié

lexique	TP	précision	rappel	f-mesure
<i>seeds</i>	98	0.859	0.388	0.535
<i>pos</i>	120	0.869	0.476	0.615
<i>wikidata</i>	100	0.862	0.396	0.543
<i>jdm morpho</i>	119	0.856	0.472	0.608
<i>jdm</i>	132	0.469	0.523	0.495
<i>brown</i>	105	0.541	0.416	0.470
<i>w2v seeds</i>	109	0.480	0.432	0.455
<i>w2v misspell</i>	102	0.698	0.404	0.512

Objectif

Ressources morphologiques

Ressources sémantiques

Méthodes distributionnelles

Meta

Meta

vote : termes présents dans au moins 2 lexiques différents

total : termes présents dans tous les lexiques

lexique	TP	précision	rappel	f-mesure
<i>seeds</i>	98	0.859	0.388	0.535
<i>pos</i>	120	0.869	0.476	0.615
<i>wikidata</i>	100	0.862	0.396	0.543
<i>jdm morpho</i>	119	0.856	0.472	0.608
<i>jdm</i>	132	0.469	0.523	0.495
<i>brown</i>	105	0.541	0.416	0.470
<i>w2v seeds</i>	109	0.480	0.432	0.455
<i>w2v misspell</i>	102	0.698	0.404	0.512
<i>vote</i>	130	0.702	0.515	0.594
<i>total</i>	145	0.331	0.575	0.420



Brown, P., deSouza, P., Mercer, R., Della Pietra, V., and Lai, J. (1992).
Class-based n-gram models of natural language.

Computational Linguistics, 18(4):467–479.



Eholié, S., Tapi Nzali, M. D., Bringay, S., and Jonquet, C. (2016).

MuEVo, a breast cancer Consumer Health Vocabulary built out of web forums.

In Paschke, A., Burger, A., Splendiani, A., Marshall, M., and Romano, P., editors, *SWAT4LS: Semantic Web Applications and Tools for Life Sciences*, Amsterdam, Netherlands.



Grabar, N. and Hamon, T. (2014).

Automatic extraction of layman names for technical medical terms.

In *ICHI 2014*, Pavia, Italy.







Lafourcade, M. (2007).

Making people play for lexical acquisition.

7th Symposium on Natural Language Processing.

jeuxdemots.

-  Liang, P. (2005).
Semi-Supervised Learning for Natural Language.
Master, Massachusetts Institute of Technology, Boston, USA.
-  Messai, R., Zeng, Q., Mousseau, M., and Simonet, M. (2006).
Building a bilingual french-english patient-oriented terminology for breast cancer.
In *MedNet*.
-  Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a).
Efficient estimation of word representations in vector space.
In *Workshop at ICLR*.
-  Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b).
Distributed representations of words and phrases and their compositionality.
In *NIPS*.



Schmid, H. (1994).

Probabilistic part-of-speech tagging using decision trees.
In *ICNMLP*, pages 44–49, Manchester, UK.
treetagger.



Zeng, Q. and Tse, T. (2006).

Exploring and developing consumer health vocabularies.
JAMIA, 13:24–29.