Drug-related questions classification

Matthieu Cordonnier and Samuel Hurault

ENS Data Challenge : Posos

April, 3 2019

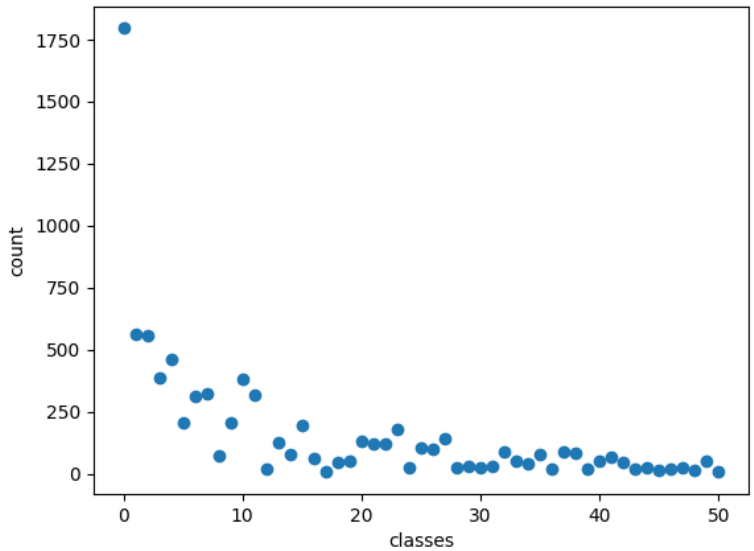
Medical questions classification

51 classes :

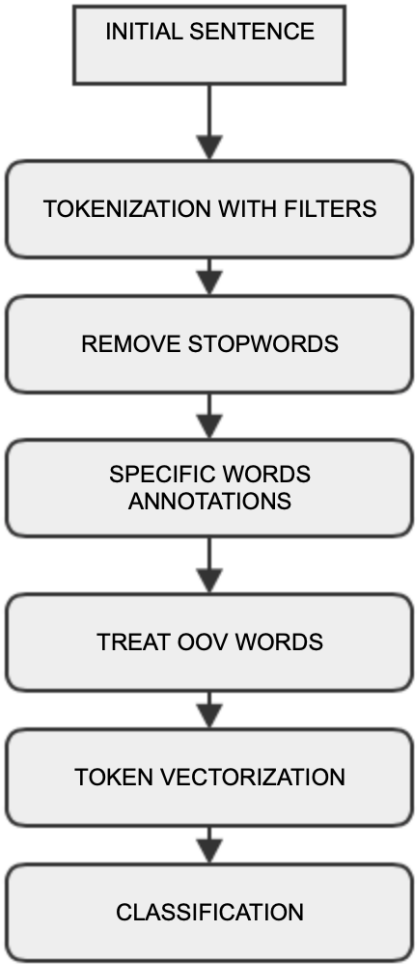
- "mon psy me dit de prendre 50mg de sertraline le matin et 50 mg de sertraline le soir. peut on prendre 100mg soit le matin ou a midi?" $\longrightarrow$ Posologie
- "est ce que la pilule trimordiol fais grossir?" $\longrightarrow$ Effets secondaires

Main difficulties :

- Spelling errors and abbreviations
  " je suis en train de prendre une pillule qui s'ppelle daily gé. "
- Technical medical vocabulary
  " oubli androcur, conseils ? "
- Low amount of data : 8000 training sentences for 51 classes (IMDB movie review : 25 000 for 2 classe) .
- Unbalanced dataset



Our approach



We mainly focused on data prepossessing

Data preprocessing

- Remove stopwords
- Medical annotations
- Lemmatization

patch evra commencé à j+4 donc est ce que je ne suis pas protégée pdt 1 mois?

↓

'patch', 'evra', 'commencé', 'à', 'j', 'donc', 'est', 'ce', 'que', 'je', 'ne', 'suis', 'pas', 'protégée', 'pdt', 'mois'

Remove Stopwords

Stop word : Very common in the language, useless for classification task → "le", "la", "ce" .

'patch', 'evra', 'commencé', 'à', 'j', 'donc', 'est', 'ce', 'que', 'je', 'ne', 'suis', 'pas', 'protégée', 'pdt', 'mois'

↓

'patch', 'evra', 'commencé', 'donc', 'protégée', 'pdt', 'mois'

Annotations

Annotation : Replacing words from a common lexical field by a single word. QUAERO Medical corpus[1] and VIDAL data base [2].

- medicine $\leftrightarrow$ "médicament".
- human anatomy $\leftrightarrow$ "*anatomie*".

'patch', **'evra'**, 'commencé', 'à', 'j', 'donc', 'est', 'ce', 'que', 'je', 'ne', 'suis', 'pas', 'protégée', 'pdt', 'mois'

$\downarrow$

'patch', **'médicament'**, 'commencé', 'donc', 'protégée', 'pdt', 'mois'

[1] Aurélie Névéol et al. "The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization". In: *Proc of BioTextMining Work*. 2014, pp. 24–30.

[2] `https://www.vidal.fr/Sommaires/Medicaments-A.htm`

Lemmatizing

Lemmatizing : grouping together the inflected forms of a word.
'patch', 'médicament', 'commencé', 'donc', **'protégée'**, 'pdt', 'mois'

↓

'patch', 'médicament', 'commencé', 'donc', **'protégé'**, 'pdt', 'mois'

Embedding

Embedding : token $w \mapsto$ sequence of vector of dimension d.

Low amount of data $\longrightarrow$ use a pretrained embedding dictionary : 150000 wod trained on the Ffrwac corpus.

Out-Of-Vocabulary module

Embedding dictionary : dictionary of reference from French pretrained embedding[3].
OOV : a word in our data not in the embedding dictionary.

	number of OOVs
Initial sentence	3209
Annotate meds	2283
Annotate anatomy	2272

Table: Number of OOVs during preprocessing

[3] Jean-Philippe Fauconnier. *French Word Embeddings*. 2015. URL: http://fauconnier.github.io.

OOV correction

Damerau–Levenshtein (DL) distance measures formal distance between strings

If a word is not in the embedding dictionary

- List all the words in **the embedding dictionary** at maximum DL distance 2 from the input word.
- return the word that appear the most in the training corpus.
- If no word at DL distance $\leqslant 2$, don't treat the word.

Nb of OOV words remaining : 65

Data augmentation

Random synonym replacement :

- Wolf corpus[4] ("French Wordnet") gives a list of possible synonyms for each word.
- Select the most similar synonym in Embedding space.

For each word :

- with proba p : replace by best synonym.
- with proba $1 - p$: keep the word.

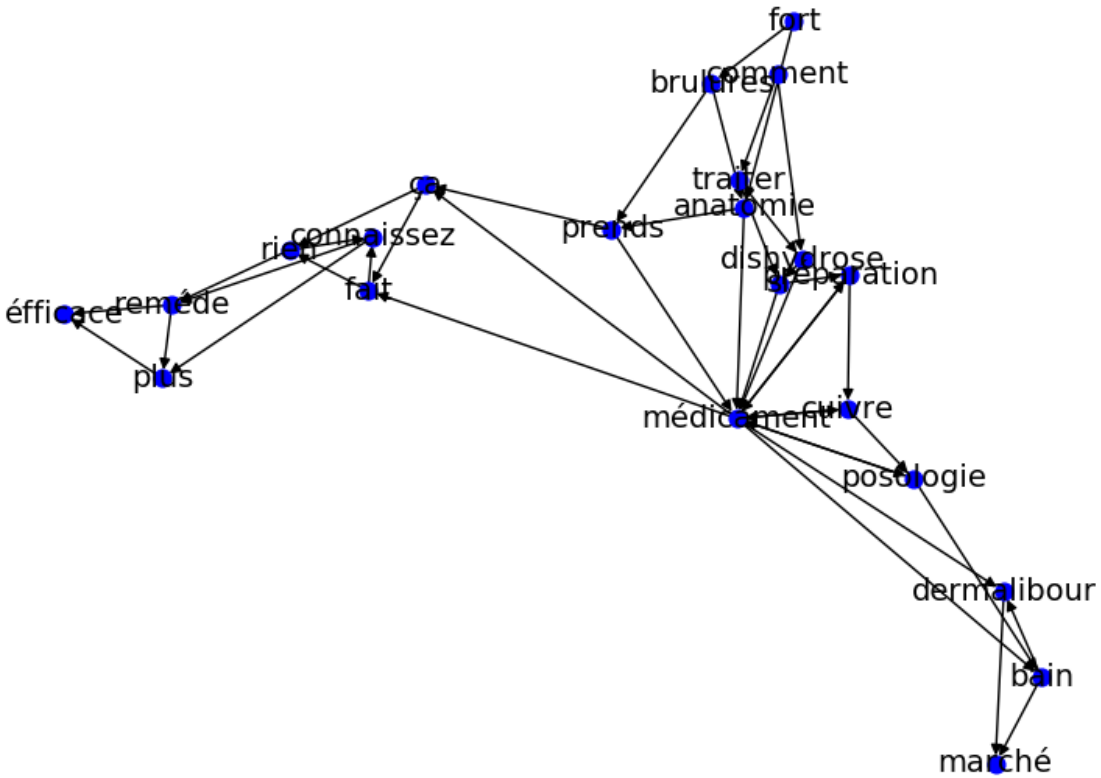
[4] Sagot Benoît et Fišer Darja. "Building a free French wordnet from multilingual resources.". In: *Ontole*. 2008.

Synonym replacement example with $p = 0.5$

- 'médicament', 'nausée', 'saignement' $\longrightarrow$ 'médicament', 'maladie', 'hémorragie'
- 'laroxyl', 'dose', 'faible', 'stress' $\longrightarrow$ 'laroxyl', 'dosage', 'faible', 'anxiété'
- 'quel', 'médicament', 'contre', 'cancer', 'anatomie', 'anatomie' $\longrightarrow$ 'quel', 'médicament', 'contre', 'sein', 'anatomie', 'anatomie'

Random walk in graph of words

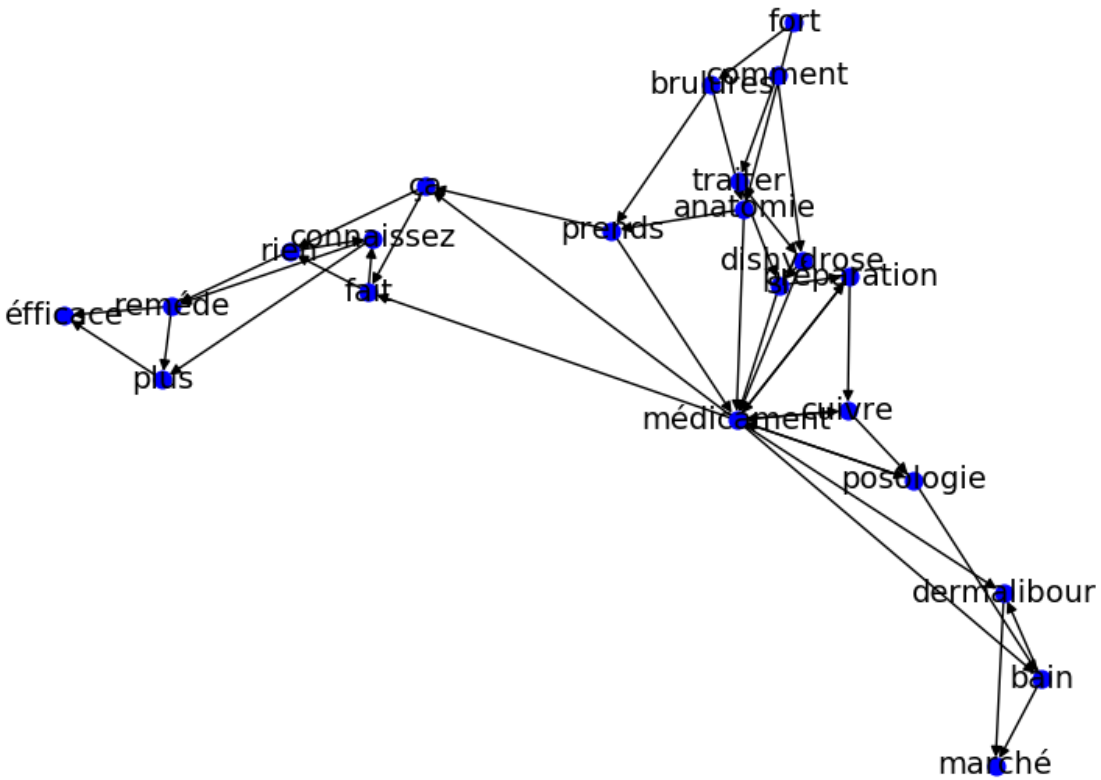
- **nodes** : words
- **weighted directed edges** : directed co-occurrence within a sliding window of size 3.



Random walk in graph of words

Random walks :

- Initialized randomly
- Transition probabilities : normalized directed weights
- Stops if no neighbors or the maximum length *maxlen* reached.



Random walk example

"Med origins" class with 7 training samples. 5 random walks with $maxlen = 10$:

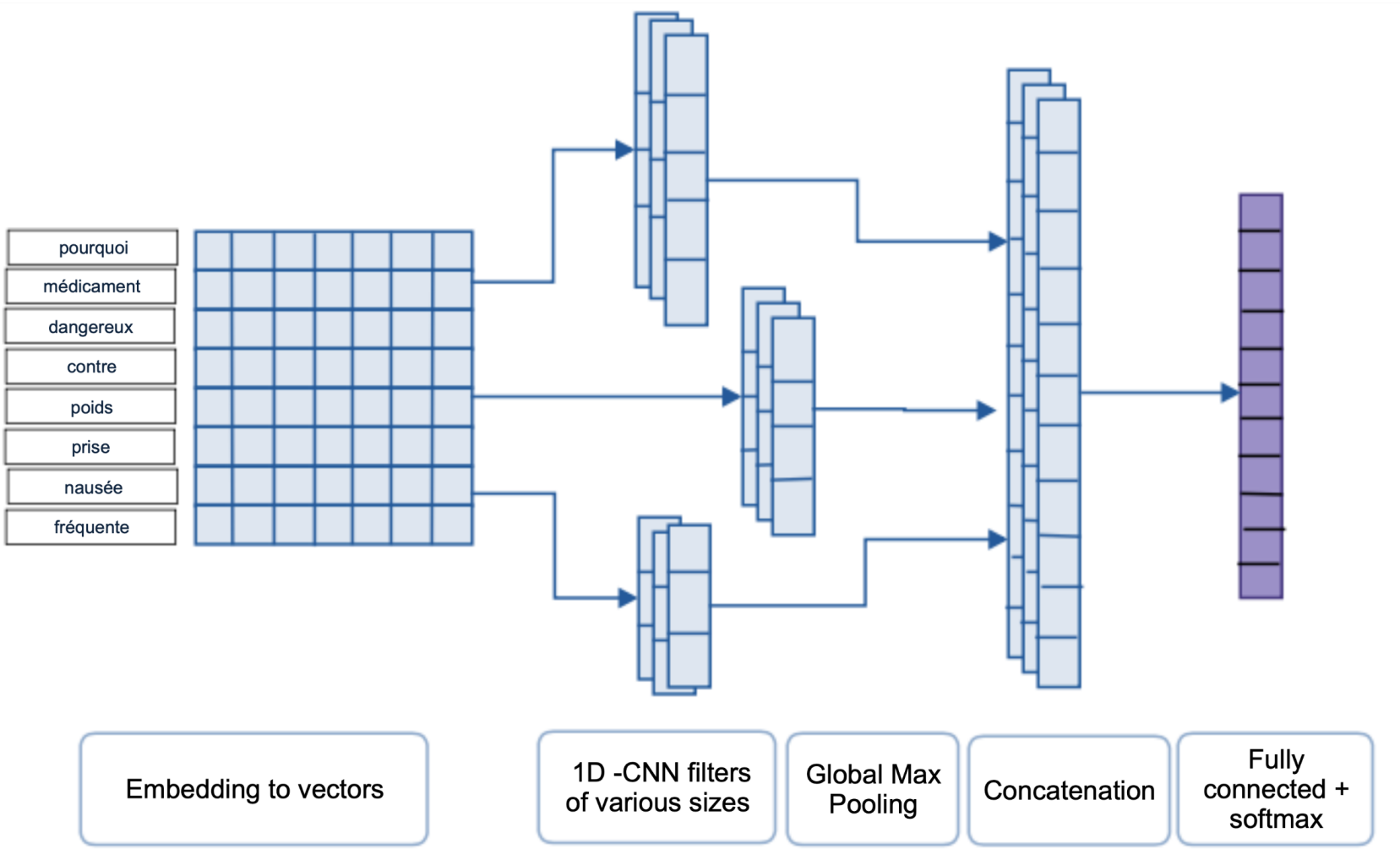
- 'vient', 'principe', 'médicament', 'origine', 'excipient'
- 'connaître', 'origine', 'animale', 'excipient'
- 'souhaite', 'connaître', 'origine', 'indigotine', 'gélule', 'médicament', 'origine', 'présent', 'lyoc', 'médicament"
- 'pénicillineet', 'voudrais', 'vrai', 'médicament', 'naturel', 'vrai', 'fait', 'anatomie', 'jument'
- 'naturel', 'vrai', 'fait', 'anatomie', 'jument'

Data augmentation parameters

Parameter	Parameter space	Optimal value
p	[0.,0.1,0.2,0.3]	0
maxlen	[0,10,20,50,100]	20
nb_per_class	[0,50,100,200]	50

Table: Data augmentation parameters tuning by cross-validation

1D Convolunational network



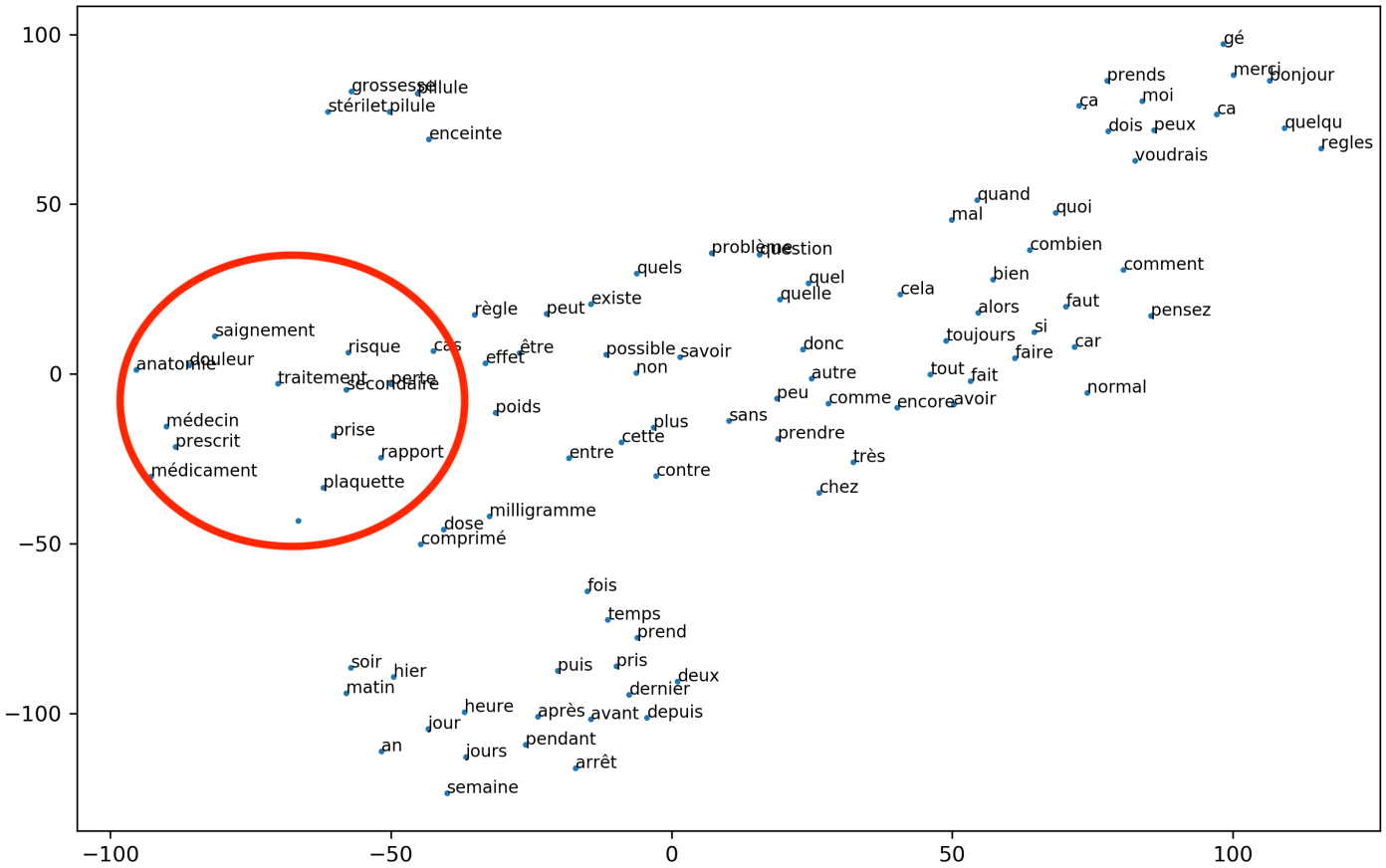
Embedding Layer

	Accuracy
No pretrained embedding	0.657
Pretrained embedding "trainable"	0.638
Pretrained embedding "not trainable"	0.435

Table: Accuracy with and without pretrained embedding initialization

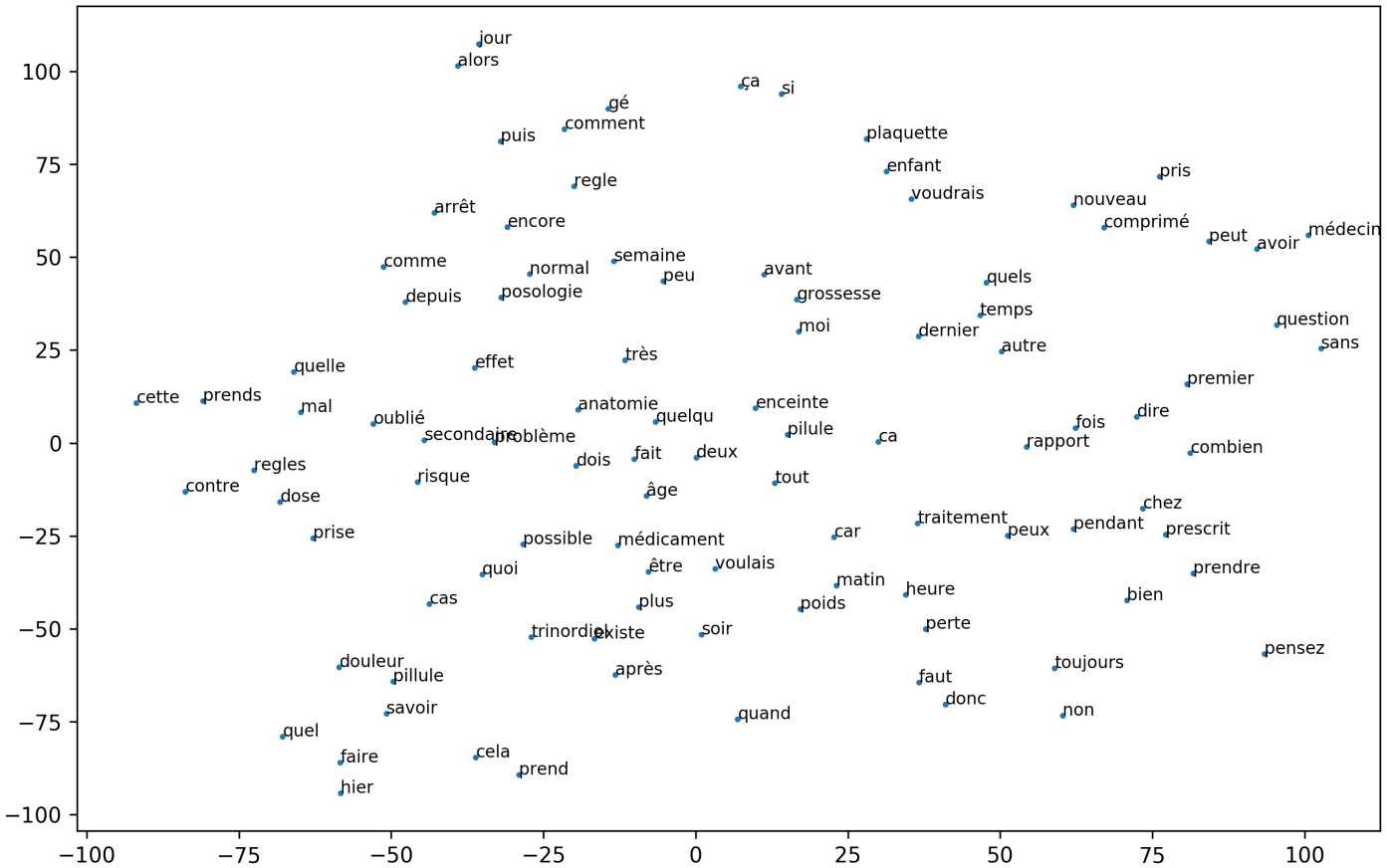
Pretrained embedding visualization

t-SNE visualization of 100 most frequent word embeddings



Trained embedding layer visualization

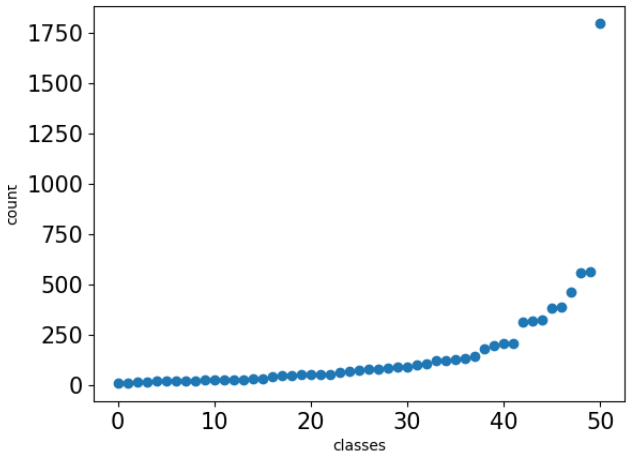
t-SNE visualization of 100 most frequent word embeddings

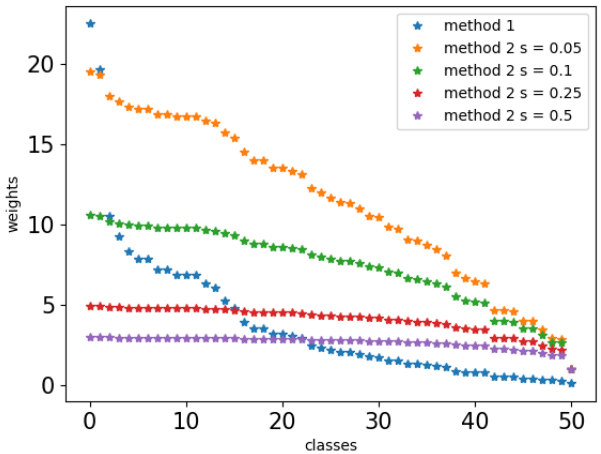


Weighted loss for class imbalance

$$L(y, x) = -\sum_{j \in C} w_j log(s_j)$$

- method 1 : $w_j = \frac{N}{N_j * nb\_class}$
- method 2 : $w_j = \frac{2 * N_{max}}{N_j + s * N_{max}}$





Parameters tuning

We tune by cross-correlation the following parameters :

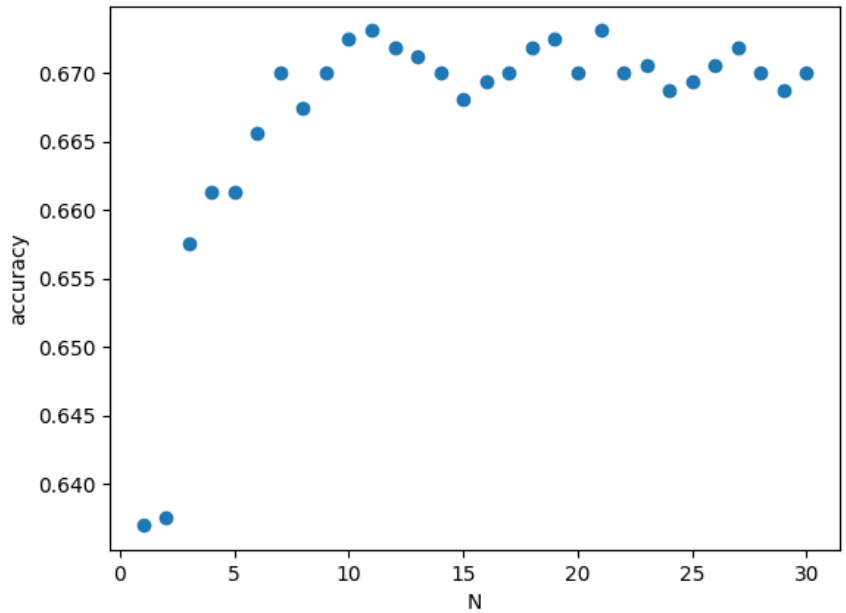
- l the lenght of the sentences in the input.
- s the weights of the loss function.

Parameter	**Parameter space**	**Optimal value**
l	$\{50, 100, 150, 200\}$	150
loss function	{normal, method 1, method 2}	method 2
s	$\{0.05, 0.1, 0.25\}$	0.25

Table: Model hyper-parameters tuning by cross-validation

Bagging method

- Train N models on 90% of data randomly sampled
- Aggregate predictions with voting



- Choice : $N = 10$

Conclusion

- Mainly focused on data preprocessing.
- Tried Attention LSTM networks without success.
- Need to train a specific medical embedding / Character-level embedding ?
- Disappointing results, fail to avoid overffiting.