ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

When was this sentence writen?

bertrand.gaiffe@atilf.fr

may 25th, 2021



www.atilf.fr

cnrs

UNIVERSITÉ DE LORRAINE

The Context

Some of us are interested in older states of the french language :

- DMF (dictionnary of middle french)
- Frantext...
- a whole team in the lab (historical linguistics of french and romance languages)

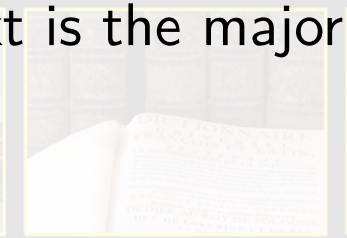
Therefore, why not a BERT model for (in particular) middle french?

- a group of people (B. Crabbé, M. Constant, L. Grobol, P. Ortiz, T. Mickus...)
- a corpus of the form (DATE SENTENCE)+

The corpus

- Frantext (texts before 1920)
- charters
- Ambroise Paré, Champlain, etc.
- OEF (Ovide en Français livre 1)
- Ovide moralisé en prose

Frantext is the major part....


Why not guessing the date?

Because we have this corpus, why not building a date guesser ?
So the problem is to guess the date when a "sentence" was written.
Examples :

- - Attendez donc, reprit le conteur ; voici le plus beau et le plus terrible de l'histoire que je vous avais promise : Coupe-en-Deux avait tombé par terre comme un plomb ; il était si soûl, si soûl, qu'il ne remuait pas plus qu'une bûche..
- - Renart, dist Nobles, bien as dit, ja de ce n'i avras desdit: or diras, nos t'escoterons; se tu diz bien, nos nos tairons.
- Helene Vous estes fou, Mendoce. Dom Diegue Vous estes folle, Helene, avecque vostre nopce.

Why not guessing the date?

Because we have this corpus, why not building a date guesser ?
So the problem is to guess the date when a "sentence" was written.
Examples :

- - Attendez donc, reprit le conteur ; voici le plus beau et le plus terrible de l'histoire que je vous avais promise : Coupe-en-Deux avait tombé par terre comme un plomb ; il était si soûl, si soûl, qu'il ne remuait pas plus qu'une bûche.. (1843)
- - Renart, dist Nobles, bien as dit, ja de ce n'i avras desdit: or diras, nos t'escoterons; se tu diz bien, nos nos tairons.
- Helene Vous estes fou, Mendoce. Dom Diegue Vous estes folle, Helene, avecque vostre nopce.

Why not guessing the date?

Because we have this corpus, why not building a date guesser ?
So the problem is to guess the date when a "sentence" was written.
Examples :

- - Attendez donc, reprit le conteur ; voici le plus beau et le plus terrible de l'histoire que je vous avais promise : Coupe-en-Deux avait tombé par terre comme un plomb ; il était si soûl, si soûl, qu'il ne remuait pas plus qu'une bûche.. (1843)
- - Renart, dist Nobles, bien as dit, ja de ce n'i avras desdit: or diras, nos t'escoterons; se tu diz bien, nos nos tairons. (1190)
- Helene Vous estes fou, Mendoce. Dom Diegue Vous estes folle, Helene, avecque vostre nopce.

Why not guessing the date?

Because we have this corpus, why not building a date guesser ?
So the problem is to guess the date when a "sentence" was written.
Examples :

- - Attendez donc, reprit le conteur ; voici le plus beau et le plus terrible de l'histoire que je vous avais promise : Coupe-en-Deux avait tombé par terre comme un plomb ; il était si soûl, si soûl, qu'il ne remuait pas plus qu'une bûche.. (1843)
- - Renart, dist Nobles, bien as dit, ja de ce n'i avras desdit: or diras, nos t'escoterons; se tu diz bien, nos nos tairons. (1190)
- Helene Vous estes fou, Mendoce. Dom Diegue Vous estes folle, Helene, avecque vostre nopce. (1650)

Characteristics of the corpus

A large part of the corpus is from Frantext. Therefore, a lot of XIXth century sentences.

Most basic baselines :

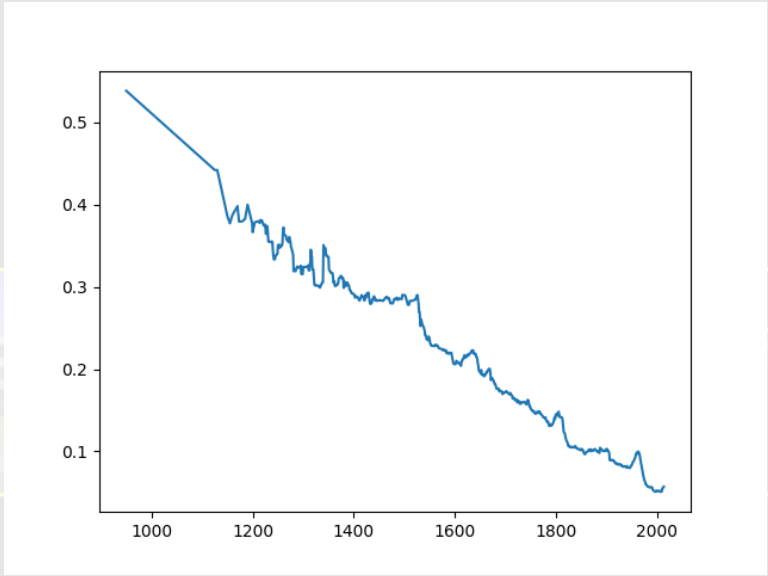
- most frequent year (on train) : 1910 (average error (on test) = 107.75)
- mean (in train) : 1805 (average error (on test) = 96.44)

An experiment about trigrams

Total variation distance (date 1, date 2):

$$\sum_{tri} |P(tri/d1) - P(tri/d2)|$$

(because we pondered wether fine tuning an existing BERT...)



tvd(d, 2000) smoothed on 20 years

Trigrams and Bayes rule

- Compute $P(Trigramme/Date)$ for all trigrams
- and also $P(trigramme)$ and $P(date)$
- then $P(date/trigramme) = \frac{P(trigrame/date) \times P(date)}{P(trigramme)}$

and then, we do something absolutely wrong (because no independance), that is :

$$P(date/t1, t2, ..., tn) = \frac{P(t1/date) \times P(t2/date).... \times P(date)}{P(t1) \times P(t2)...}$$

Entirelly false, but we are interested in argMax for the dates anyway....

Results with trigrams

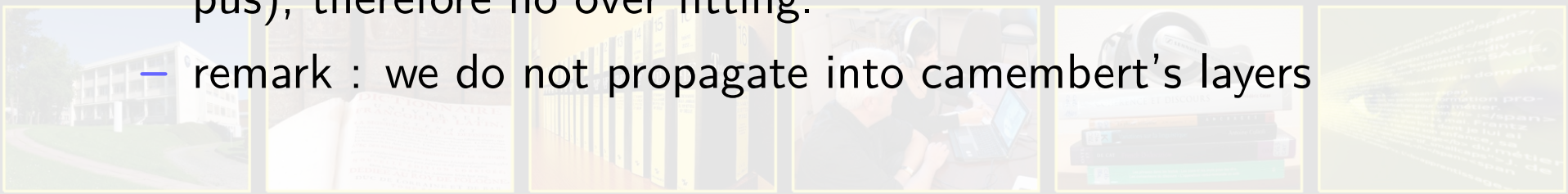
- We compute (count) the probas on TRAIN and we test on TEST.
- Result : mean error = 31 years ! (31.47)

Remember : mean date (1805) gives an average error of 96 years

Using BERT for datation

Experiment :

- Take some of the layers, and add a linear classifier on them, take the mean, then relu, tanh and normalize between 900 and 1920
- Remark, we consider dates as numbers (not categories)
- we do not even learn a full epoch (just around 5% of the corpus), therefore no over fitting.
- remark : we do not propagate into camembert's layers



results

- layer n° 12 $\longrightarrow$ 68 years
- layer n° 1 $\longrightarrow$ 98 years

These results correspond to about $1/20^{\text{th}}$ of an epoch. Therefore, no overfitting !

Is it a wordpieces problem ?

The Camembert model relies on wordpieces (not trigrams).
Could that be the problem ?
Well, same strategy as trigrams but with wordpieces gives 35 years error in average. (as compared to 31).

Very partial conclusions

- should compute not only mean error but variance (and Bert model shoul be better on this)
- should smooth the $P(tri/date)$ in the trigram approach.
- test with several layers (see wether different informations) (first test : layer 12 + layer 9 not better than layer 12 alone.