Using TreeTagger trained with modern-language data on John Walker's 18th century dictionary

Francois **HUANG**
Blanche **MIRET**
Preethi **SRINIVASAN**
Dao **THAUVIN**
Univ Paris Diderot – Sorbonne Paris Cité

Ce travail est l'œuvre conjointe d'étudiants de la Licence Informatique et de la Licence d'Études Anglophones de Paris Diderot. Il a été financièrement supporté par le programme IdEx Université de Paris ANR-18-IDEX-0001

Introduction

Since its development by Helmut Schmid in 1994, the TreeTagger has successfully been used to tag various languages. The neural network on which the tool is based can be trained with some tagged data, then used to annotate a same language corpus with POS tags and lemma information. This research tends to analyse how the TreeTagger trained with modern-language data reacts when executed on an 18th dictionary corpus. In particular, how would it respond when confronted with obsolete words? For instance we can find in the 18th century texts, words such as *advenes, affusing, gayety* which are no longer currently used; would it be able to correctly tag these words? More broadly, what information could the tool give about the evolution of a language over centuries?

Data and methods

The research is based on **John Walker's Critical Pronouncing Dictionary**(1), working on an XML file(2) representing the dictionary with machine-readable format.

INDIVIDUALLY, în-dè-vîd'-ù-âl-è, *ad.* With separate or distinct existence, numerically.
To INDIVIDUATE, în-dè-vîd'ù-âte, *v. a.* To distinguish from others of the same species, to make single.
INDIVIDUATION, în-dè-vîd-ù-â'shûn, *s.* That which makes an individual.
INDIVIDUITY, în-dè-vîd-ù'è-tè, *s.* The state of being an individual, separate existence.
INDIVISIBILITY, în-dè-vîz-è-bîl'è-tè, } 552.
INDIVISIBLENESS, în-dè-vîz'è-bl-nès, }
3 r

INDOLENT, în'dò-lênt, *a.* Free from pain; careless, lazy, inattentive, listless.
INDOLENTLY, în'dò-lênt-lè, *ad.* With freedom from pain; carelessly, lazily, inattentively, listlessly.
To INDOW, în-dôû', *v. a.* To portion, to enrich with gifts.—See *Endow*.
INDRAUGHT, în'drâft, *s.* An opening in the land, into which the sea flows; inlet, passage inwards.
To INDRENCH, în-drênsh', *v. a.* To soak, to drown.
INDUBIOUS, în-dû'bè-ûs, *a.* Not doubtful, not suspecting, certain.

Extract of the *Critical Pronouncing Dictionary* (1824)

Procedure :

- **Adapt data** to TreeTagger-expected structure via some Python scripts

- **Run** the TreeTagger

Hindsight on the result :

- The **recall** rate indicates for each POS category how well it is **recognized**, again on a 200 words sample :

$$\frac{\text{\# rightly identified words}}{\text{\# words of the same POS category}}$$

- The **precision** rate indicates the **accuracy** of the tagging calculated in the following way :

$$\frac{\text{\# rightly tagged words}}{\text{\# total words}}$$

Looking at words marked as **"unknown"** lemma is the lead to distinguish obsolete words.

Obsolete words :

A word is determined obsolete in four ways, according to the Oxford English Dictionary:

- No dictionary entries found for the word

- Frequency less or equal to $\frac{3}{8}$

- Obsolete

- No frequency shown

Discussion about the results

Global precision is lower than official rate, which could signify some degree of language evolution.

The accuracy of the "unknown" lemmas to identify obsolete words is promising on the sample. However, 89,0% of the identified "unknown" lemmas concerned out-of context words, which are all the words defined in the dictionary, introducing some blurred area about the interpretation.

Remarks for extensive analysis :

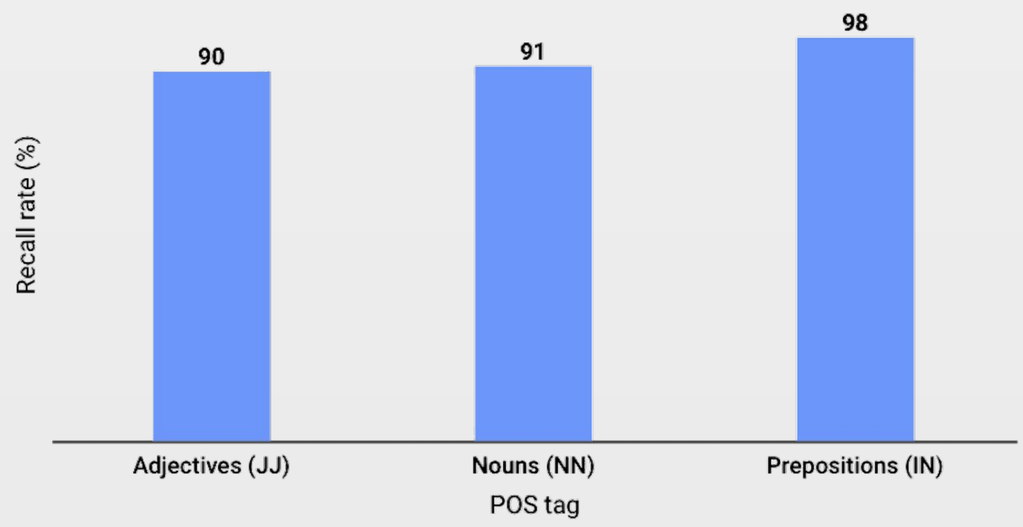
- The out-of-context words could be separately analyzed.
- The sample could be larger and taken in different parts of the dictionary.
- The obsolescence criteria could be disputed.

Results

Word	Treetagger's POS tag	Lemma
With	IN	with
separate	JJ	separate
or	CC	or
distinct	JJ	distinct
existence	NN	existence
,	,	,
numerically	RB	numerically
.	SENT	.

Precision rate of the result : **93,5%.**
Official rate of the TreeTagger on Penn-TreeBank data : **96,36%**(3)

Recall rate per POS category



Obsolete words :

Total **words** in the data : **482 240**, including :

- 98 068 punctuation signs

- 46 067 out-of-context words

Total **unknown identified** lemmas : 35 748, including :

- 31 813 out-of-context words

Accuracy of using 'unknown' marked lemma as obsolete indicator : 58,4%.

Word	Treetagger's POS tag	Lemma
understeward	NN	unknown
zodiack	NN	unknown
unploughed	JJ	unknown
shortnecked	JJ	unknown
balsamick	NN	unknown
bathingroom	NN	unknown
trafficks	NNS	unknown

Conclusion

Out of a sample of about 200 words that were assigned "unknown" as lemma, 58,4% were found to be really words that are no longer used today, such as "absterging", "denison", or "calumniator". Without being very specific, it seems that using TreeTagger trained by ancient data can be a way to measure the evolution of a language through the centuries.

References

1. Walker J. (1824) *A Critical Pronouncing Dictionary & Expositor of The English Language*

2. *Courtesy* of N. Trapateau

3. Schmid, H. (1994). *Probabilistic part-of-speech tagging using decision trees. In International Conference on New Methods in Language Processing*, pages 44-49, Manchester, UK.

4. D. Khurana, A. Koli, K. Khatter, S. Singh. *Natural Language Processing: State of The Art, Current Trends and Challenges*, Manav Rachma International University, 2017.