Comparing John Walker's 18th century grammatical categories against those of today with Treetagger

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

Introduction

As it is very well known, Treetagger (*Helmut Schmid*, 1996) is a tool that is widely used in the Natural Language Processing domain to annotate multilingual corpora with POS taggers and lemma information. The purpose of this study is to find out whether Treetagger could learn the grammatical categories of John Walker's *Critical Pronouncing Dictionary* (1791), and how it will react by tagging 20th century's corpora.

Data

The data used for this investigation are the Walker's Dictionary, used to produce a lexicon and a list of tags, a part of the Brown Corpus used to train treetagger and test .

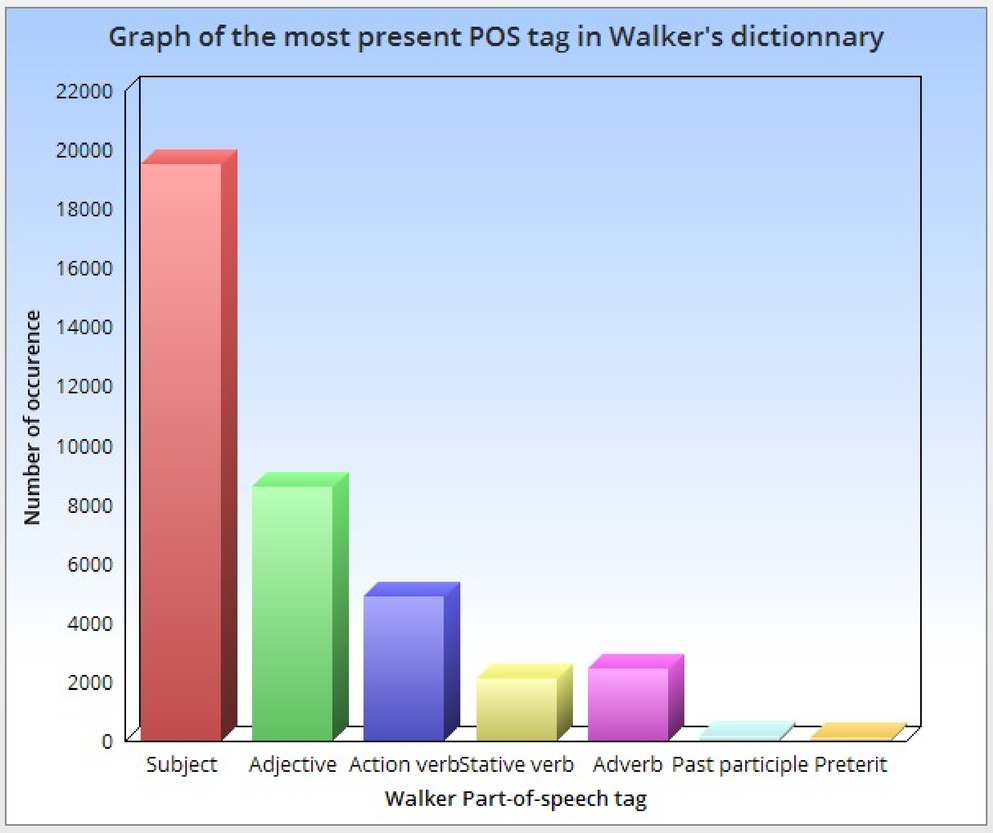
First Steps

Brown Corpus's tagset and grammatical categories of Walker's dictionary don't match because many categories changed or simply disappeared over the course of time. In other words, Brown Corpus's tags of 20th century and Walker's tags of 18th century are not always the same, it could possibly influence the result's accuracy. Some Walker's dictionary's words were not tagged. 37895 words were tagged and 879 were not. We deleted the untagged words from our lexicon.

Token	Walker's grammatical categories	Treetagger's POS tags
Abacus	s.	NN
Beneficed	a.	VBN
To Fight	v. a.	VB
Foreworn	part.	JJ
To can	v.a.	VB

Remarks on Walker's Dictionary

Over 38 000 words, these are the most used POS by Walker.



Walker used many tags as Part-Of-Speech tags but the Brown Corpus doesn't consider all of them: for e.g. there are no *plural, contraction* or *A negative or privative termination.*

The word *MEN* is tagged as **plural** but the word *WOMEN* isn't.

The tag **contraction** hasn't been used for every words which required it. For instance, the word *NE'ER* is a poetical contraction for Never and the word *TA'EN* is a poetical contraction for Taken. While *NE'ER* is declared as an adverb, *TA'EN* is declared as a contraction.

Experiment

We mapped the Walker's Dictionary's grammatical categories with Brown Corpus tagset, the Brown Corpus's tags not used in the mapping were replaced by other tags of the Brown Corpus. We applied this mapping on a part of the Brown Corpus. We trained treetagger with that part of the Brown Corpus. We tagged another part of the Brown Corpus with the obtained tagger and compared tags given by Brown Corpus and our tagger.

Preparing TreeTagger

About our lexicon :

We used the Walker's dictionary definitions to get the tagged words, we added some proper nouns, punctuation marks and cardinal numbers from our training set.

Notes :

- Walker's dictionary uses different tags for the same category.
- Some categories have disappeared, such as *solemn nominative plural.* (Hapax: Ye), and *A negative or privative termination.* (Hapax: Less).

About our training set :

We borrowed a piece of the Brown corpus to train tree-tagger with about 1500 sentences and 28639 words (including punctuation marks).

Treetagger's POS tags	Brown Corpus's POS tags
art	at, dt
adj	jj, dti, ap, dts, jjt, jjr, jjs
pret	vbd, dod, bed, bedz, hvd
s	nn, nr, nns, nps
conj	in, cs, cc, dtx
prep	in, to
part.pass.	hvd, vbn, md
tp (third person)	vbz, bez, hvz, doz
v.	bem, vb, hv, be, do, ber
adv.	rb, ql, abx, abn, abl, ex
pron.	pps, ppo, pn, wps, ppl, wdt
part.	vbg
interj.	uh

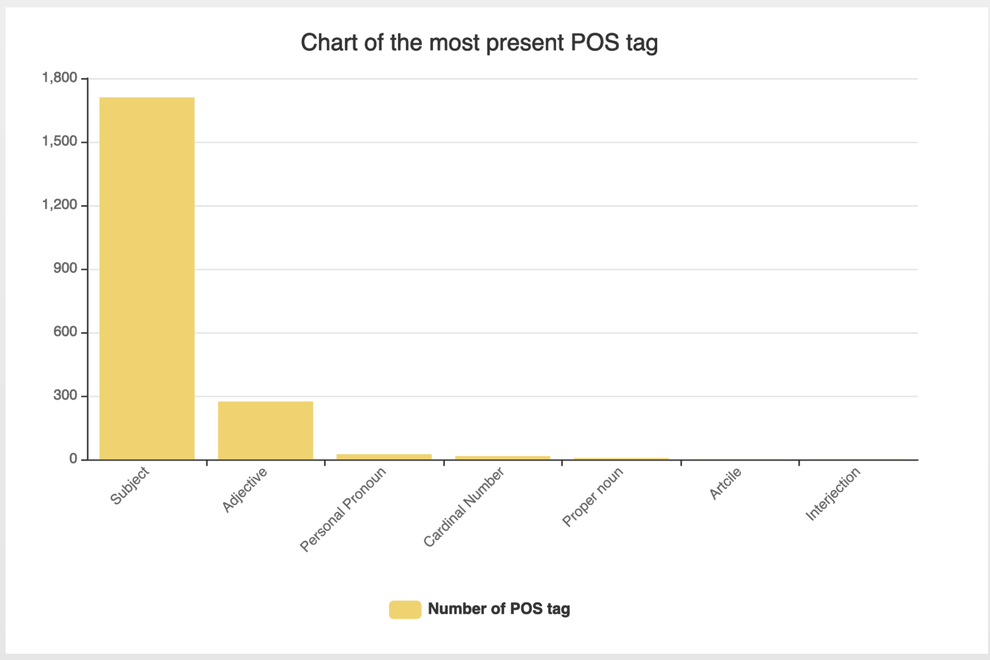
Results

We used our tree-tagger to tag a brown corpus text and compare with the tags given by the Brown corpus.This is what we can observe :

1590 failures on 2332 tags.

A big amount of s. with 1300 failures in our tagging.

Only 8 tags have been used : adj., s., SENT, pron.pers., CD, NP, art., interj.



Discussion about the results

The tagset we obtained with our mapping has only 19 tags, compared to the Brown Corpus tagset size of 86 tags. Therefore, our small number of tags probably impacts the result. Training on a bigger training set could give a better result. Most of the tags not used are rare tags that don't often appear in the training set. For example, the tag sp (for second person) appears only five times.
Our test set has only 2332 words (with punctuations), it is surely not enough to have an accurate result. In our case, even the number of failures is too much for such a small amount of test data.

References

1. J. Walker. *A Critical Pronouncing Dictionary*, 1824.
2. N. Trapateau. *A Critical Pronouncing Dictionary* from J. Walker xml file
3. W. N. Francis and H. Kucera. *Brown Corpus*
4. The project's Github : https://github.com/daothauvin/TreeTaggerWithWalker