EVALUATING A MODERN POLISH POS-TAGGER ON HISTORICAL DATA

A project by

Maria Irena Szawerna for LT2314 Language Technology Resources

Background

◦ POS information can be useful in other types of research (e.g. corpus linguistics).

◦ POS taggers are created for big modern languages or require much data to train.

◦ Solution: using modern POS taggers on historical data?

◦ Issues: OOV items, variation in spelling and punctuation, different grammar…

◦ These problems can be solved, but first they need to be identified.

◦ Most research has been done on Germanic languages.

RESEARCH QUESTION:

How well does the GATE Cloud tagger for Polish perform on historical data and what kinds of mistakes does it tend to make?

Materials and methods

◦ GATE Cloud tagger for Polish, trained on Universal Dependencies treebanks.

◦ Manually annotated data from 19th century memoirs, 3000+ tokens.

◦ Code in a Jupyter Notebook file: data preprocessing, tagger annotation, an array of evaluation measures (quantitative and qualitative).

◦ Manual error analysis.

RESULTS

POS tag	Capitalized	Lowercase
ADJ	**61.51%**	52.56%
ADP	95.65%	**95.94%**
ADV	**72.29%**	70.11%
AUX	**76.92%**	73.91%
CCONJ	**97.52%**	97.50%
DET	98.46%	**98.67%**
NOUN	**84.94%**	73.23%
NUM	**73.68%**	69.44%
PART	82.22%	**82.61%**
PRON	82.89%	**83.12%**
PROPN	**66.23%**	0.00%
PUNCT	100.00%	100.00%
SCONJ	**80.43%**	77.50%
VERB	**84.40%**	82.45%
X	**64.52%**	64.00%

Table 2: Precision per POS tag per trial.

POS tag	Capitalized	Lowercase
ADJ	**78.90%**	78.06%
ADP	97.06%	**97.35%**
ADV	70.59%	**71.76%**
AUX	74.63%	**76.12%**
CCONJ	**95.73%**	95.12%
DET	41.11%	**48.68%**
NOUN	79.56%	**89.54%**
NUM	**65.12%**	58.14%
PART	59.68%	**61.29%**
PRON	88.73%	**92.66%**
PROPN	**93.09%**	0.00%
PUNCT	100.00%	100.00%
SCONJ	**53.62%**	44.93%
VERB	90.38%	90.38%
X	**52.63%**	42.11%

Table 3: Recall per POS tag per trial.

Data capitalization	Accuracy	MCC
Original capitalization	83.98%	81.82%
All lowercase	80.46%	77.80%

Table 1: Accuracy and MCC per trial.

Confusion matrix (original capitalization)
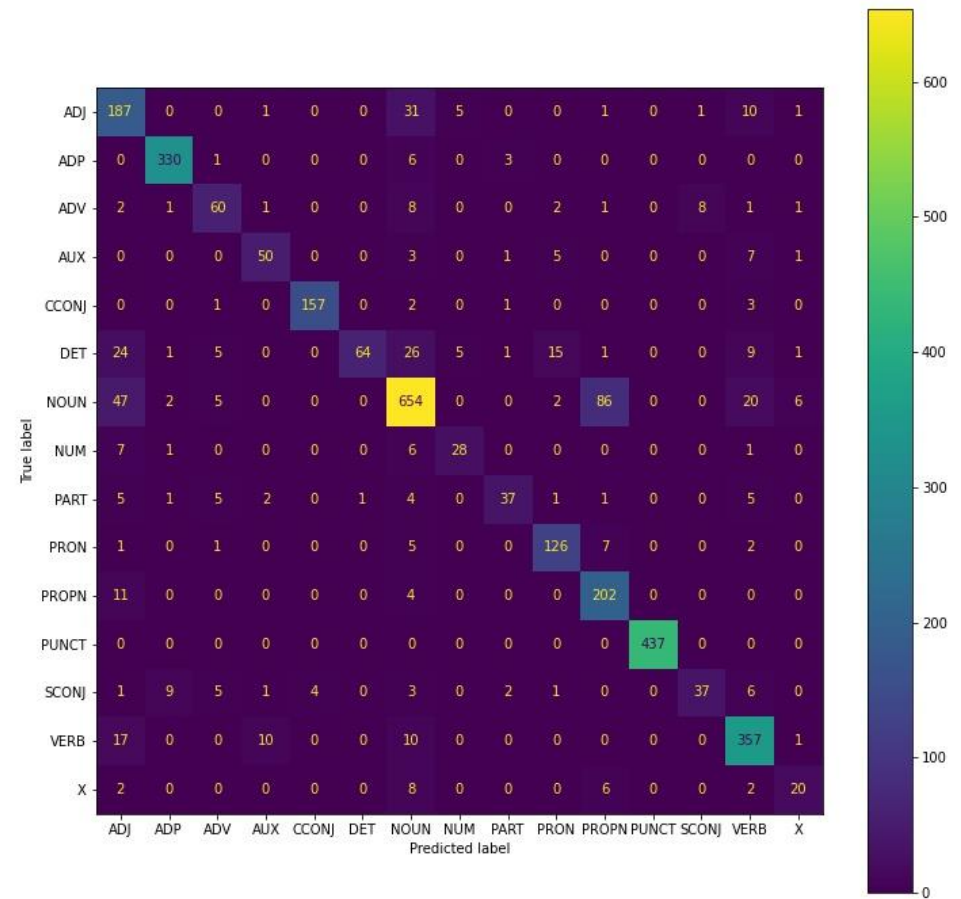
Confusion matrix (lowercase)
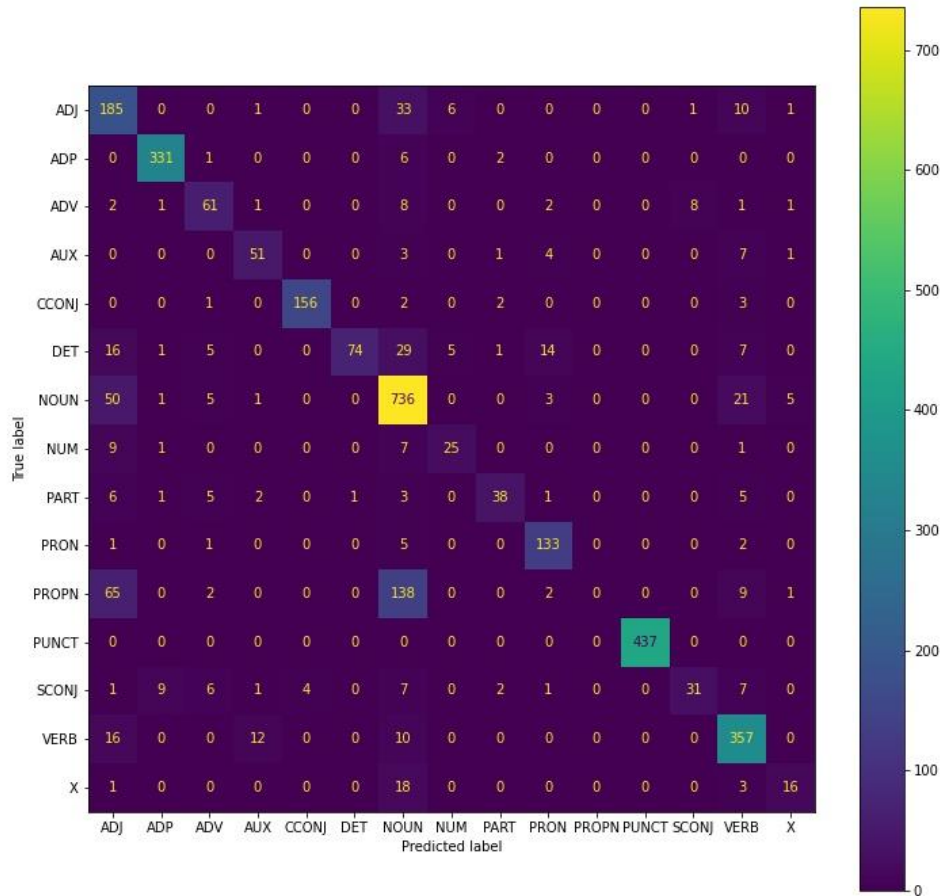
Qualitative error analysis

- [Jupyter Notebook](#)

- Capitalization, confusing word endings, variation in spelling, the ability of some words to function as multiple classes, and archaic vocabulary.

- Mysterious cases.

Errors: examples (original capitalization)

○ Word-final *y* instead of *j*:

333	szczupłey	tak szczupłey dotacyi	ADJ	VERB	0.495477

○ Ambiguous word class:

508	1830	albo 1830 roku	ADJ	NUM	0.868293

○ Archaic or nonstandard spelling:

243	wedle	poddani wedle ilości	ADP	NOUN	0.963168

○ Nonstandard capitalization:

424	Dziadek	bo Dziadek móy	NOUN	PROPN	0.793743

○ Surnames (adjectival paradigm):

114	Czermińskiej	starościny Czermińskiej we	PROPN	ADJ	0.875157

○ UD annotation-induced class ambiguity:

449	Był	. Był bardzo	VERB	AUX	0.876343

○ Lexical/grammatical similarity to Polish:

408	Mychayłowu	mojomu Mychayłowu .	X	PROPN	0.829579

Discussion

◦ Performance better than expected (above 80%).

◦ Issues similar to the ones described in the literature, except the confusing endings.

◦ Lowercasing the data does not bring obvious benefits, and is detrimental to PROPN detection.

Future Work

◦ Influence of punctuation.

◦ Testing other taggers on the same data.

◦ Testing this tagger on different kinds of historical data.

◦ Testing taggers on modern nonstandard data.

◦ Developing methods for preprocessing the data.

THANK YOU FOR YOUR ATTENTION