CMC Tagger in the Task of Tagging Historical Texts

Wiktor Walentynowicz & Tomasz Kot

Task

- Morphological disambiguation and segmentation of Polish texts for different century.

- Input: a graph of tokens representing paragraphs.

- Output: a tagged graph of tokens representing paragraphs.

Zasięg	Segment	Lemat	Znacznik
0–1	|Miał|	mieć	praet:sg:m1.m2.m3:imperf
1–2	|em|	być	aglt:sg:pri:imperf:wok
0–2	|Miałem|	miał	subst:sg:inst:m3
2–3	|dużo|	dużo:d	adv:pos
		dużo:n	num:sg.pl:nom.gen.acc:m1.m2.m3.f.n:rec
3–4	|jabłek|	jabłko	subst:pl:gen:n:ncol
4–5	|.|	.	interp

Idea

Take standard architecture for tagging.

Extend it for segmentation.

Tagger - Input Vector

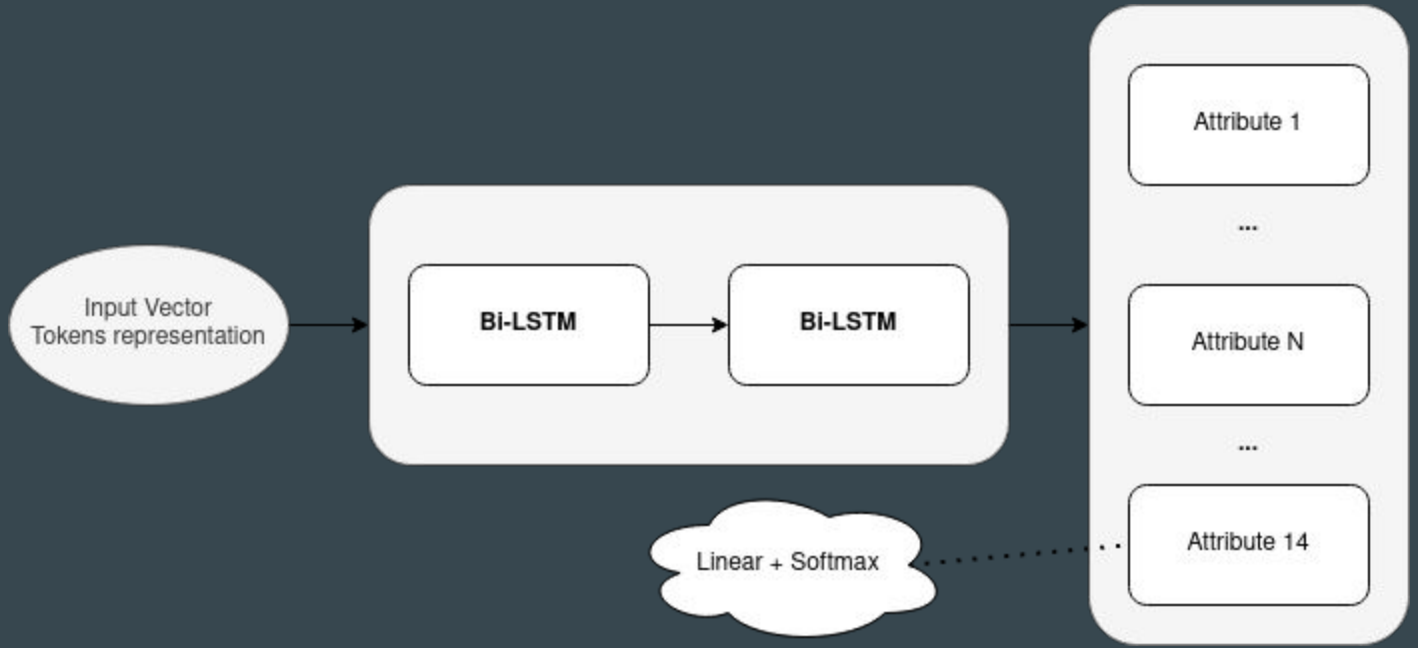
Morphological information vector
(binary)

Suffix Character Embedding Vector
(floating point)

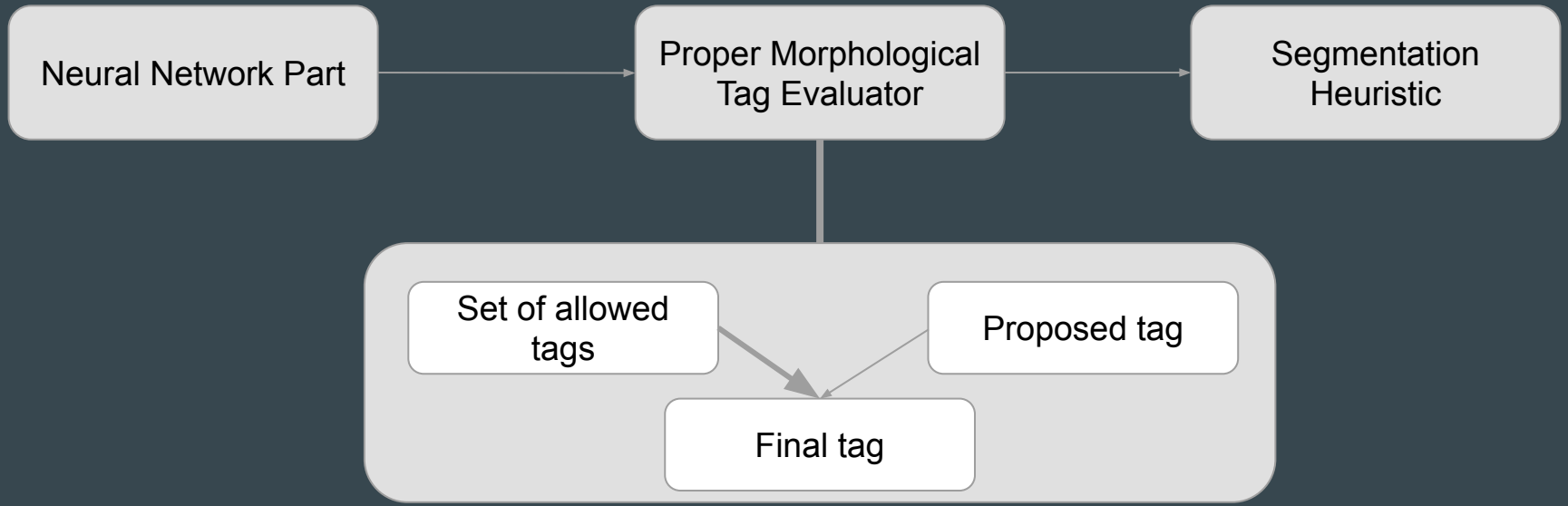
fastText embedding vector
(floating point)

Brown Cluster representation vector
(floating point)

Tagger - Neural Network Architecture



Tagger - Whole System



Segmentation Heuristics

- Always shortest path
- Always longest path
- Statistical approach for ambigousion path

Experiments - Tagger for each Text Epoch and General Tagger

Short Path Heuristic

Measure	17	ALL_17	19	ALL_19	20	ALL_20	COMB	ALL_COMB
Overall	82.96%	84.03%	92.19%	92.18%	92.81%	93.75%	87.70%	88.42%
Known	84.15%	85.07%	92.81%	92.76%	93.47%	94.38%	88.64%	89.27%
Unknown	40.84%	47.07%	49.41%	51.18%	49.59%	52.07%	43.85%	52.07%
Manual	29.97%	33.20%	41.67%	40.26%	38.96%	40.26%	33.31%	40.26%

Long Path Heuristic

Measure	17	ALL_17	19	ALL_19	20	ALL_20	COMB	ALL_COMB
Overall	87.59%	88.68%	92.89%	92.94%	92.91%	*93.88%*	90.25%	91.00%
Known	88.90%	89.85%	93.52%	93.54%	93.57%	*94.52%*	91.24%	91.91%
Unknown	40.84%	47.07%	49.41%	51.18%	49.59%	*52.07%*	43.85%	48.63%
Manual	30.37%	33.50%	41.67%	44.17%	38.96%	*40.26%*	33.58%	40.26%

Experiments - Fine Tuning for specific Epoch

Measure	Short 17	Long 17	Short 19	Long 19	Short 20	Long 20
Overall	84.10%	*88.83%*	92.82%	*93.62%*	93.47%	93.63%
Known	85.14%	*90.00%*	93.41%	*94.22%*	94.09%	94.28%
Unknown	46.89%	*46.89%*	51.76%	*51.76%*	52.89%	52.89%
Manual	33.10%	*33.60%*	44.17%	*44.17%*	40.69%	40.69%

Final version

For 17th century texts TL Model
For 19th century texts TL Model
For 20th century texts All Text Model
All on Long Heuristic

Measure	Final Multitagger
Overall	91.21%
Known	92.14%
Unknown	50.72%
Manual	16.70%

https://gitlab.clarin-pl.eu/syntactic-tools/morphological/cmc-tagger/-/tree/cmc-heuristics

Conclusions