Stanza, spaCy ja UDPipe'i eesti keele mudelite võrdlus

Stanza

Stanza süntaksimudelite kasutamiseks on vaja eelnevalt sõnestust ja lausestust, lemmatiseerimist ning POS-märgendust. Stanza mudelite kasutamise kohta EstNLTK teegis leiab juhendi siit.

StanfordNLP Stanza puhul on vajalikud töövoo komponendid (sõnestuseks, lemmatiseerimiseks jen) kättesaadavad ja lihtsasti kasutatavad Stanza teegi kaudu. StanfordNLP Stanza eesti keele mudel on integreeritud ka EstNLTK teeki. Sel puhul kasutatakse EstNLTK tokenisatsiooni ning StanfordNLP lemmatiseerijat ja POS-taggerit.

Stanza + morph_analysis on kasutatav EstNLTK töövoos StanzaSyntaxTaggeri abil. Parsimiseks on vaja, et tekst oleks sõnestatud ja lausestatud ning sisaldaks Vabamorfi morfoloogilist analüüsi. Otse Stanza teegis saab mudelit kasutada, kui järgida eeltokeniseeritud/märgendatud dokumentide kasutamise juhist, milles väljad täita vastavate Vabamorfi märgenditega, kusjuures *feats* peab olema kujul *sg=sg|all=all* ja UPOS ning XPOS sisaldavad sama Vabamorfi sõnaliigimärgendit.

Stanza + morph_extended on analoogne Stanza + morph_analysis mudeliga, kuid kasutab täiendatud Vabamorfi märgendust.

Spacy

SpaCy mudelid on spaCy töövoos kasutatavad eraldiseisva komponendina, st ette saab anda tokeniseerimata, lemmatiseerimata ja muu töötluseta teksti. Infot töövoo seadistamise kohta leiab spaCy lehelt. SpaCy eesti keele mudelite installimisest moodulitena ning nende kasutamisest parsimisel leiab infot GitHubist.

Spacy CPU vähem täpne, kui transformereid sisaldavad mudelid, kuid erinevalt viimastest on see efektiivne ka CPUd kasutades.

Spacy + EstBERT on TartuNLP/EstBERT transformeri põhine mudel. Soovitatav on kasutada GPUd.

Spacy + RoBERTa-xlm on RoBERTa-xlm mitmekeelse transformeri põhine mudel. Soovitatav on kasutada GPUd.

UDPipe

UDPipe'i mudelid on integreeritud EstNLTK teeki. Mudelite kasutamiseks on vajalik sõnestus, lausestus ja Vabamorfi morfoloogline analüüs. Kasutamise juhise EstNLTKs leiab siit.

UDPipe + morph_analysis on kasutatav EstNLTK teegi kaudu ning vajab *morph_analysis* analüüsikihti.

UDPipe + morph_extended on kasutatav EstNLTK teegi kaudu ning vajab *morph_extended* analüüsikihti.

	eeltöötluseta (töövoost sõltumatult) kasutatav	Integreeritud EstNLTKs	Vabamorfi märgendusel põhinev	LAS	UAS
StanfordNLP Stanza*		✓		83.81	86.69
Stanza + morph_analysis		✓	✓	86.13	88.44
Stanza + morph_extended		✓	✓	**86.30**	**88.64**
Spacy CPU*	✓			62.71	70.79
Spacy + EstBERT*	✓			83.47	86.21
Spacy + RoBERTa-xlm*	✓			85.49	88.08
UDPipe + morph_analysis		✓	✓	73.90	78.94
UDPipe + morph_extended		✓	✓	76.63	80.85

Table 1: Mudelite võrdlus. * Mudel hinnatud automaatse sõnestuse ja lausestuse põhjal.