

Součástí úkolu jsou Python3 skripty `assignment3_1.py` a `assignment3_2.py`.

Task 1

Použili jsme Brillův tagger z knihovny NLTK. Jako počáteční tagger jsme použili jednoduchý unigramový tagger, který si zapamatává nejčastější značku pro každé slovo trénovacích dat a pro neznámé slova vrátí speciální značku “None”. Použili jsme počáteční množinu 24 vzorů¹ (template), které původně použil Brill. Pro angličtinu jsme počet pravidel obmedzili² na 247, pro češtinu na 500. Tieto konkrétne hodnoty sme dostali skúšaním niekoľkých hodnôt. Všeobecne platí, že väčšie množstvo pravidiel môže mierne zlepšiť presnosť, ale výrazne spomaňuje.

Výsledky

Angličtina	Přesnost	Čeština	Přesnost
1. stupeň	0.8909342163585638	1. stupeň	0.7743503905627291
2. stupeň	0.8890974014126724	2. stupeň	0.7887507306445614
3. stupeň	0.8747591522157996	3. stupeň	0.7618758256274769
4. stupeň	0.8935028395769291	4. stupeň	0.7919649224755012
5. stupeň	0.8925100950892275	5. stupeň	0.7880596223212392
Průměr	0.888160740931	Průměr	0.781000298326
σ	0.00686479392893	σ	0.0113145302014
Průměr triv.	0.865374314735	Průměr triv.	0.738888267001
σ triv.	0.0143254860843	σ triv.	0.0134321134672

Tabulka 1: Přesnost (accuracy) pro Brillův tagger pro oba jazyky.

¹NLTK má tieto vzory ako súčasť knižnice

²Limit sa pri týchto počtoch vždy naplní, čiže limit je rovnaký ako počet pravidiel.

Task 2

Supervizovaný HMM

Metóda Přesnost	Čas
Žiadny pruning, add λ 0.9252470471671056	7h 24m 22s
Nevyhladený OOV model 0.9268375355252523	1m 34s
Vyhladený OOV model 0.9280369201887727	5h 35m 13s
OOV a obmedzené stavy 0.927802257971997	1h 12m 24s
Add λ a obmedzené stavy 0.9250123849503298	1h 23m 32s
OOV model, pruning 10 0.9279847730294892	1m 53s
OOV model, pruning 10 0.9280369201887727	3m 45s

Tabulka 2: Přesnost (accuracy) pro Brillův tagger pro oba jazyky.

Angličtina	Přesnost	Čeština	Přesnost
1. stupeň	0.9280369201887727	1. stupeň	0.791274775492853
2. stupeň	0.9227200458727552	2. stupeň	0.8096870184388119
3. stupeň	0.9277601208774032	3. stupeň	0.7636195508586526
4. stupeň	0.9170442118418997	4. stupeň	0.8034285110541747
5. stupeň	0.93424514784421	5. stupeň	0.8035752174835561
Průměr	0.92596128932500821	Průměr	0.80233748927646664
σ	0.00576517615271536	σ	0.0060165626151515

Tabulka 3: Přesnost pro oba jazyky určené cross-validací pro model z prořezávacím limitem 20. Pro češtinu se omezejeme pouze na stavy, které jsou v trénovacích datech, pro angličtinu používáme jako stavy všechny dvojice značek.

Nesupervizovaný HMM — Baum-Welch

Pre túto časť sme sa rozhodli použiť ako stavy iba bigramy, ktoré sme videli v trénovacích dátach, pretože inak by bol počet stavov bol veľký.

Navyše sme sa rozhodli použiť pruningovú metódu, aby sme tréning urýchlili.