Jana Novotná & Vladan Glončák
Statistical NLP II
Assignemnt 3

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

Task 1

Použili sme Brillov tagger z knižnice NLTK. Ako počiatočný tagger sme použili jednoduchý unigramový tagger, ktorý si zapamätá najčastejšie značku pre každé slovo trénovacích dat a pre neznáme slová vráti špecialnú značku "None". Použili sme počiatočnú množinu 24 vzorov[1] (template), ktoré pôvodne použil Brill. Pre angličtinu sme početpravidiel obmedzili[2] na 247, pre 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**
1. stupeň	0.8909342163585638
2. stupeň	0.8890974014126724
3. stupeň	0.8747591522157996
4. stupeň	0.8935028395769291
5. stupeň	0.8925100950892275
Průměr	0.888160740931
σ	0.00686479392893
Průmer triv.	0.865374314735
σ triv.	0.0143254860843

Čeština	**Přesnost**
1. stupeň	0.7743503905627291
2. stupeň	0.7887507306445614
3. stupeň	0.7618758256274769
4. stupeň	0.7919649224755012
5. stupeň	0.7880596223212392
Průměr	0.781000298326
σ	0.0113145302014
Průmer triv.	0.738888267001
σ triv.	0.0134321134672

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

[1] NLTK má tieto vzory ako súčasť knižnice
[2] Limit sa pri týchto počtoch vždy naplní, čiže limit je rovnaký ako počet pravidiel.

Task 2

Supervizovaný HMM

Angličtina	**Přesnost**
1. stupeň	0.9280369201887727
2. stupeň	0.9227200458727552
3. stupeň	0.9277601208774032
4. stupeň	0.9170442118418997
5. stupeň	0.93424514784421
Průměr	0.92596128932500821
σ	0.00576517615271536

Čeština	**Přesnost**
1. stupeň	0.791274775492853
2. stupeň	0.8096870184388119
3. stupeň	0.7636195508586526
4. stupeň	0.8034285110541747
5. stupeň	0.8035752174835561
Průměr	0.80233748927646664
σ	0.0060165626151515

Tabulka 2: Přesnost pro oba jazyky určené cross-validací pro model z prořezávacím limitem 20. Pro češtinu se omezejume 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.