Stanford POS Tagger

Lubomír Sedlář, 359 719

4. prosince 2012

Stanford Loglinear Part-Of-Speech Tagger je nástroj pro přiřazování morfologických značek jednotlivým slovům textu. Jde o implementaci taggeru popsaného v článku [1] v jazyce Java.

Program vyžaduje k běhu Javu ve verzi alespoň 1.6. Ke stažení je dostupných více verzí programu. Základní verze obsahuje pouze dva naučené modely pro angličtinu, plná verze obsahuje modely i pro další jazyky, jako arabština, čínština nebo němčina.

Pro angličtinu tagger ve výstupu používá klasické značky z Penn Treebank Popis značek pro ostatní jazyky je dostupný ve stáhnutém balíku.

Tagger je licencován pod GNU General Public License verze 2 nebo pozdější, je tedy k dispozici kompletní zdrojový kód. Balík obsahuje programy umožňující spouštění jako program na příkazové řádce nebo jako server.

Existují také další rozšíření, ať už grafická rozhraní nebo API pro další jazyky jako Ruby nebo Python.

Implementační detaily

Učební algoritmus taggeru je založen na maximální entropii (*Maximum entropy classifier*). Tento model je podobný naivnímu bayesovskému klasifikátoru, učení je ovšem pomalejší. Model se snaží určit logaritmus pravděpodobnosti jevu jako $\log Pr(Y_i = k) = \beta_k \cdot X_i - \ln Z$ pro $k \in \{1, \ldots, K\}$, kde X_i je vektor proměnných popisující pozorovaný jev i a β_k je vektor vah odpovídající očekávanému výsledku k. $\ln Z$ zajišťuje, aby tato množina pravděpodobností tvořila pravděpodobnostní rozložení, tedy měla součet 1. Hodnotu Z lze vypočítat na základě vektorů β a X.

Tagger přiřazuje slovu w značku t na základě kontextu h, což je typicky několik předcházejících slov a jejich značky. Základní verze modelu používá pro klasifikaci slova w_i ve větě $w_1 \ldots w_n$ se značkami $t_1 \ldots t_n$ kontext $\{t_{i-2}, t_{i-1}, w_{i-1}, w_i\}$, tedy značky dvou předcházejících slov, aktuální a předchozí slovo.

Ke zpřesnění značkování podstatných se používá řada pomocných heuristik: jeden dodatečný atribut je informace o tom, jestli je slovo psané kompletně velkými písmeny. Taková slova mají výrazně jiné rozložení značek než odpovídá slovům, ve kterých se pouze vyskytuje nějaké velké písmeno. Další atribut určuje, jestli slovo začíná velkým písmenem, ale přitom se nenachází na začátku větu.

Pro zpřesnění značkování sloves se používá větší kontext – tagger se dívá na až 8 předcházejících slov (zastavuje se, pokud narazí na sloveso) a hledá slova jako *to*, modální slovesa nebo slova, která se často pojí s infinitivem (*make*, *help*, aj.).

Pro částice se používá doménová znalost. Pokud je značkované slovo na seznamu slov často tvořících částice, hledá se v kontextu sloveso, o kterém víme, že se často pojí se značkovaným slovem jako částicí. Tato doménová znalost byla sestavena analýzou učících dat ve fázi předzpracování.

Výsledky

V balíku dostupném ke stažení je několik naučených modelů pro angličtinu. Nejpřesnější z nich správně klasifikuje 97,28 % slov. Na dříve neviděných slovech má správnost 90,46 %. Tento model používá nejen levý kontext, ale dívá se i na dosud neoznačkovaná slova.

Protože nepřesnější model je poměrně pomalý, je doporučeno použít model používající pouze levý kontext. Ten dosahuje správnosti 96,97 % (88,85 % na neznámých slovech). Doporučovaný model dokáže označkovat na „2008 nothing-special Intel server“ až 15000 slov za sekundu.

Modely pro čínštinu mají přesnost 93–94 %, pro arabštinu 96,5 % a pro němčinu 96,9 %. Přesnost modelů pro francouzštinu není uvedena.

Zdroje

[1] Kristina Toutanova and Christopher D. Manning. 2000. *Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger.* In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63-70.