TiMBL – Shallow Parser

Tilburg Memory-Based Learner (TiMBL) je open-source softwarový nástroj vyvinutý na univerzitě v nizozemském Tilburgu, který úspěšně implementuje algoritmy učení z instancí (*memory-based learning*). Nejpoužívanějšími dvěma algoritmy jsou váhovaný algoritmus k nejbližších sousedů *IB1 -IG* a *IGTree*, jeho efektivní aproximace pomocí rozhodovacího stromu.

Shallow parsing (*chunking*, „mělká“, „povrchová“ analýza) je proces určení shluků slov tvořících jmenné, předložkové a slovesné fráze, tedy hlavní konstituenty vět, a vztahů mezi nimi.

Využití strojového učení nás zbavuje nutnosti manuální definice relací, závislosti na konkrétních datech, ba i jazyku. Učení z instancí pak oproti ostatním metodám strojového učení umožňuje lépe si poradit s množstvím výjimek, které se v přirozených jazycích vyskytují, a odlišit je od šumu. Je také vhodné pro kombinaci mnoha heterogenních zdrojů informací díky využití techniky podobnostního vyhlazování (*smoothing-by-similarity*) tam, kde jsou data neúplná.

Při využití tohoto nástroje chápeme syntaktickou analýzu textu v přirozeném jazyce jako sadu klasifikačních úloh. Cílem jedné klasifikační úlohy je pro jednotlivá slova určit, zda jsou či nejsou součástí fráze (jmenné, slovesné…), další určuje vztahy mezi frázemi – například zda je jmenná fráze podmětem či předmětem fráze slovesné. Proces analýzy pak zahrnuje sekvenci takovýchto klasifikačních úloh, kde výstup jedné slouží jako vstup další (tento kaskádový přístup se objevuje i v řadě metod analýzy textu založených na jiných principech).

Úlohou chunkingu je přiřadit každému slovu jednu ze značek označující jeho příslušnost do určitého typu fráze: máme značky „není součástí žádné fráze“ a dvojici značek pro každý typ fráze, prosté „je součástí fráze (daného typu)“ a „je prvním slovem (nové) fráze (daného typu)“. Jako vstupní informace slouží pro každé slovo lemma a slovní druh (*POS tag*) dvou předcházejících, aktuálního a následujícího slova. Na anglickém korpusu WSJ, při testování na $\frac{1}{25}$ a tréninku na zbytku korpusu TiMBL průměrně dosáhl accuracy 98 % na jmenných a 99 % na slovesných frázích.

Při hledání vztahů frází klasifikujeme typ vztahu každé potenciální dvojice frází. Jako vstup uvažujeme lemma a značku slovesa, kontext hlavy jmenné fráze obsahující dvě předchozí slova, hlavu samotnou, a následující slovo či zbytek fráze, vzdálenost mezi frázemi, kde se fráze počítají jako jediné slovo, počet mezilehlých čárek a počet mezilehlých slovesných frází (tyto numerické atributy umí správně interpretovat pouze algoritmus IB1-IG). Přidání dalších informací o slovech a frázích ke zlepšení nevede. Je důležité správně určit, jak vzdálené dvojice frází má klasifikátor zvažovat; příliš velká vzdálenost rychle vede k zahlcení šumem a přílišnému nárůstu času výpočtu. Praxe ukazuje, že v angličtině je vhodné ignorovat dvojice, mezi kterými leží více než jedna slovesná fráze. Vzhledem k tomu, že IGTree se v této úloze projevuje jako opatrnější (vyšší precision) a lépe určuje předmětné fráze, zatímco IB1-IG se chová přesně obráceně, lze nejlepších výsledků dosáhnout vhodnou hlasovací strategií. Dosažená accuracy je 97,4 %, precision 89,8 %, recall 68,6 %.

Memory-based shallow parsing je jednoduchá efektivní metoda analýzy textu pomocí strojového učení s učitelem, jejíž úspěšnost snese srovnání s konkurenčními přístupy. V principu je flexibilní a lze ji použít i na problematické úlohy (např. vnořené fráze).

TiMBL: Timburg Memory-Based Learner [on-line]. 7. října 2012 [cit. 11. 12. 2012]. WWW adresa: <http://ilk.uvt.nl/timbl/>.

DAELEMANS, Walter, BUCHHOLZ, Sabine, and VEERNSTRA, Jorn. *Memory-based Shallow Parsing*. In Proceedings of EMNLP/VLC-99, p. 239–246. University of Maryland, USA, June 1999. Dostupné on-line na <http://acl.ldc.upenn.edu/W/W99/W99-0707.pdf>.