359860 Juraj Ondruška

Stanford Named Entity Recognizer

Menná entita označuje časť textu, ktorý menom odkazuje na nejakú entitu. Môže to byť jedno, alebo aj viacslovné pomenovanie. Medzi menné entity patrí napríklad „Peter“ vo vete: „Peter číta knihu“, ale nie kniha v tej istej vete. Kniha totiž v tejto vete označuje fyzickú entitu, ale neoznačuje entitu mennú, pretože „kniha“ neurčuje meno nejakej knihy.

Rozpoznávanie menných entít (NER) je úloha spracovania textu kvôli identifikácii a klasifikácii mien v oblasti spracovania prirodzeného jazyka, ktorá poskytuje extrakciu dôležitých informácií. Rozpoznávanie menných entít sa často vykonáva pomocou štatistického taggera, ktorý sa učí vzory pre rozpoznávanie mien z ručne označených textových korpusov. NER rieši dve úlohy. Prvou je identifikácia menných entít v texte a druhou je klasifikácia entity do rôznych tried (napr. Osoba, Miesto, Čas). Jedným zo spôsobov, ako funguje NER, je klasifikácia každého slova nezávisle na sebe. Problémom tohto prístupu je, že slová nemusia byť na sebe nezávislé. Napríklad „New York“ - miesto a „New York Times“ - organizácia. Tento problém riešia tzv. „sequence models“. Stanford NER využíva linear chain Conditional Random Field sequence model. Tento model využíva diskriminatívny prístup, teda používa podmienenú pravdepodobnosť namiesto joint pravdepodobnosti. Stanfordský rozpoznávač menných entít je naprogramovaný v jazyku Java. Na identifikáciu a klasifikáciu entít používa zmieňovaný inear chain Conditional Random Field sequence model doplnený o features extraction.

Features extraction zabezbečuje extrakciu pravidiel, potrebných pre identifikáciu a klasifikáciu menných entít v texte. Features v Stanford NER využívajú poznatky z lokálneho kontextu, teda zo svojho okolia, ale tiež vedomosti o štruktúre daného slova. Z faktu, že predchádzajúce slovo dostalo zaradenie „meno osoby“ plynie veľká pravdepodobnosť, že aj nasledujúce slovo bude určené ako „meno osoby“ a podobne. Štruktúra slova tiež plní dôležitú úlohu. Záleží najmä na veľkosti písmen, prítomnosti čísel a rôznych interpunkčných znamienok, alebo tiež na konkrétnom podreťazci slova. Napríklad, ak slovo končí reťazcom „field“, tak pôjde s vysokou pravdepodobnosťou o miesto.

Stanford NER používa natrénovaný model , ale ponúka tiež možnosť vytvorenia vlastného modelu z označkovaného korpusu, alebo tiež možnosť vytvorenia vlastných „features“. Daným „features“ sú pomocou „supervised“ strojového učenia pridelené váhy a to buď negatívne, alebo pozitívne. Pri určovaní triedy, do ktorej bude slovo zaradené potom hlasuje každá „feature“, ktorej podmienky toto slovo s lokálnym kontextom spĺňuje. Trieda, ktorá dostane najviac vážených hlasov, je potom zvolená za správnu.

Stanford NER ponúka natrénované modely pre angličtinu nemčinu a čínštinu. Pre angličtinu sú to verzie, ktoré dokážu rozpoznávať rôzne skupiny menných entít. A to buď: osoba, miesto a organizácia,
osoba, miesto, organizácia a rôzne
osoba, miesto, organizácia, čas, peniaze, dátum, percentá