ROZPOZNÁVANIE POMENOVANÝCH ENTÍT V SLOVENČINE – WEBOVÉ ROZHRANIE[1]

Radovan Garabík
Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied
Bratislava

Rozpoznávanie pomenovaných entít (NER; z anglického Named-entity recognition) patrí k oblastiam počítačového spracovania prirodzeného jazyka, ktoré majú priame aplikačné použitie a sú často žiadané aj vo sférach mimo lingvistického výskumu (pri vyhľadávaní v texte alebo spracovaní textových dát). Pod pomenovanými entitami sa do veľkej miery myslia onymické objekty (aj viacdenotátové), ktoré označujú individuálny objekt (reálny alebo fiktívny), a teda tento pojem sa prekrýva s pojmom proprií. Praktické požiadavky (spočívajúce hlavne v potrebe vyhľadávania rôznorodých objektov v neštrukturovaných textových dátach, buď priamo koncovým používateľom, alebo na ďalšie spracovanie, napríklad anonymizáciu osobných či citlivých údajov) ale vedú k niektorým všeobecne zaužívaným rozšíreniam. K pomenovaným entitám sa často priraďujú aj dátumy a iné časové označenia (niekedy aj periodické, ako sú názvy mesiacov, opakujúcich sa sviatkov či dní v týždni), adresy (klasické geografické aj nové, ako napr. URL, e-mail), rôzne číselné označenia, hodnoty fyzikálnych a menových veličín a podobne.

K rozpoznávaniu pomerne často patrí aj určenie, o aký typ pomenovanej entity ide (typicky sa rozlišujú alebo sú žiadané hlavne mená osôb, geografické názvy, urbanonymá, názvy inštitúcií, časové a číselné údaje).

NER býva založené na troch základoch:

1. slovníky známych pomenovaných entít (najmä antroponymá a toponymá);
2. heuristické spracovanie entít, najmä pri rozpoznávaní URL, dátumov, časových údajov, číselných hodnôt;
3. ručne značkované korpusy (dostatočnej veľkosti), slúžiace ako základ pre trénovanie štatistických alebo iných nástrojov na NER.

Slovníky pomenovaných entít umožňujú pomerne rýchlo dosiahnuť istú základnú úroveň presnosti značkovania; je ale potrebné ich udržiavať aktualizované, entity, ktoré sa v nich nenachádzajú, nebudú rozpoznané. Slovníky neriešia problém homonymie (či už entít s ne-entitami alebo rôznych typov entít medzi sebou) či polysémie.

[1] Spracovanie pomenovaných entít je rozvíjané v rámci projektu INEA/CEF/ICT/A2019/1926831 *Curated Multilingual Language Resources for CEF AT* spolufinancovaného Európskou úniou prostredníctvom Nástroja na prepájanie Európy.

Heuristické spracovanie je obmedzené typom pomenovaných entít, ktoré sú dostatočne identifikovateľné a diferencovateľné (napr. dátumy a číselné údaje). Neveľmi dobre funguje pre typické propriá, ktorých prakticky jediným identifikátorom je veľkosť počiatočného písmena.

Vytvorenie ručne značkovaného korpusu je časovo aj personálne značne náročné[2], ak chceme dosiahnuť dostatočnú veľkosť, ale takýto korpus je dôležitý na dosiahnutie rozumnej presnosti anotácie.

Pre slovenčinu existuje niekoľko verejne dostupných webových služieb (v rôznych stupňoch funkčnosti) poskytujúcich rozpoznávanie pomenovaných entít, napr. *NER* v rámci portálu NLP Nástroje[3], *Named Entity Recognition for Slovak Language* na stránkach Fakulty informatiky a informačných technológií Slovenskej technickej univerzity[4] alebo rozpoznávanie pomenovaných entít na portáli NLP4SK: NLP as a Service[5].

Pre češtinu existuje ručne značkovaný korpus „Czech Named Entity Corpus 2.0“ s rozsahom 8 993 viet a s označenými 35 220 entitami (Straková et al., 2017). Entity sú klasifikované v dvojúrovňovej hierarchii 46 typov. Veľmi pozitívnym faktom je, že korpus je voľne dostupný (aj získateľný) pod licenciou CreativeCommons Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0).

Existencia tohto korpusu nám blízkeho jazyka bola využitá pri tvorbe rozpoznávania pomenovaných entít v slovenčine. Korpus bol strojovo preložený do slovenčiny s využitím najlepších existujúcich dostupných možností strojového prekladu[6]. Následne boli z neho odstránené vety, ktoré obsahovali nepreložené úseky textu, rozpoznané podľa prítomnosti písmen *ů, ř, ě*. Podobne boli odstránené vety, pri ktorých nebolo možné v preloženom texte identifikovať ekvivalent (preklad) pomenovanej entity. Identifikácia entity v preklade spočívala v porovnávaní redukovaných lem (bez diakritiky, bez dvojhlások, bez prípon) v originálnej a preloženej vete. Preložený korpus je dostupný na adrese https://www.juls.savba.sk/ner.html pod licenciou CC BY-NC-SA 3.0. Rozsah korpusu je po filtrácii 6 735 viet, 13 173 označených entít.

Tento preložený korpus bol použitý pri trénovaní nástroja *nametag*[7] (Straková et al. 2014), na korpuse boli natrénované dva modely na rozpoznávanie pomenovaných entít. Prvý model, nazvaný trivial, využíva iba povrchové črty textu (slovo, číselnú hodnotu, veľkosť písmen, heuristika na rozpoznanie URL a e-mailových

[2] Práce na takomto ručne značkovanom korpuse slovenčiny prebiehajú v oddelení Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra SAV

[3] http://nlp.bednarik.top/ner/

[4] http://mus.fiit.stuba.sk/

[5] http://arl6.library.sk/nlp4sk/

[6] Google Translate; https://translate.google.com

[7] https://ufal.mff.cuni.cz/nametag/1

adries) bez akéhokoľvek lingvistického spracovania. Tento model slúži primárne iba na účely porovnávania, poskytuje iba veľmi hrubé a nepresné značkovanie.

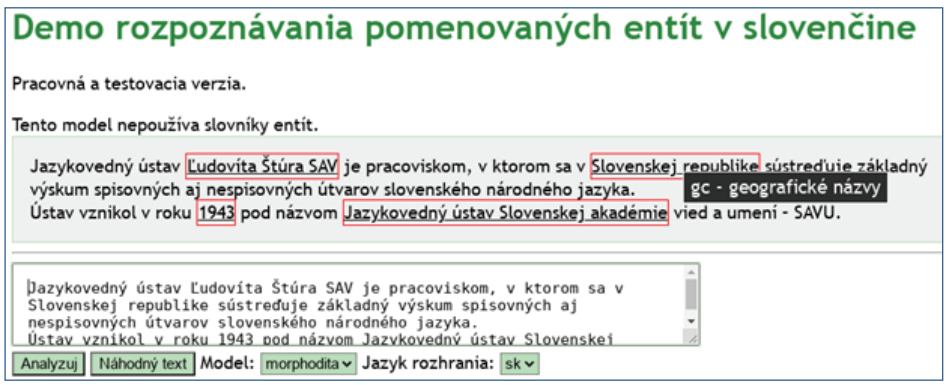
Druhý model, nazvaný *morphodita*, zahŕňa v sebe model na lematizáciu a morfologické značkovanie natrénovaný na ručne značkovanom korpuse (*r-mak-6.0*). Umožňuje tak lepšie využitie inherentnej lingvistickej informácie v texte pri rozpoznávaní entít.

Jazykovedný ústav Ľ. Štúra SAV sprístupňuje rozhranie, ktoré demonštruje tieto modely na rozpoznávanie pomenovaných entít v slovenských textoch. Rozhranie je prístupné na adrese https://www.juls.savba.sk/nerd/ (aktuálna verzia zo dňa 10. 2. 2022).

V rozhraní je možné zadať krátky text (v rozsahu niekoľkých odsekov), v ktorom budú automaticky rozpoznané pomenované entity. Okrem zadania vlastného textu je možné zadať aj voľbu *Náhodný text*, ktorá náhodne vyberie niekoľko viet slovenského textu[8] na ukážku, bez potreby zadávania vlastného textu.

Vo výsledku sú rozpoznané pomenované entity graficky zvýraznené, pričom typ entity sa zobrazí po nadídení kurzorom myši na danú entitu. Rozhranie je dostupné v slovenčine a angličtine.

Nasledujúca ukážka rozhrania obsahuje krátky text s rozpoznanými entitami, kurzor sa nachádza nad jednou z nich (*Slovenskej republike*) a je zobrazený jej typ (*geografický názov*); nasleduje textové pole určené na zadávanie textu na analýzu a ovládacie prvky – spustenie analýzy (*Analyzuj*), analýza niekoľkých náhodne vybraných viet (*Náhodný text*), výber modelu a výber jazyka rozhrania.



Opísané webové rozhranie slúži na demonštráciu slovenského modelu rozpoznávania pomenovaných entít, ktorý je dostupný verejnosti pod licenciou CC BY-NC-SA 3.0, a teda použiteľný v širokej škále možných aplikácií. Perspektívne plánujeme sprí-

[8] Výber z niekoľkých legislatívnych a informatívnych dokumentov.

stupniť aplikačné rozhranie (API), ktoré umožní model používať aj vo vzdialených aplikáciách zainteresovaných používateľov, sprístupniť ho prostredníctvom siete European Language Grid[9] a prípadne model rozšíriť o ďalšie dáta, pripravované v oddelení Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra SAV.

Literatúra

BEDNÁRIK, Filip: Extrakcia informácií z textu. Bratislava: Fakulta informatiky a informačných technológií STU v Bratislave (diplomová práca).

KAŠŠÁK, Ondrej – KOMPAN, Michal – BIELIKOVÁ, Mária: Extrakcia pomenovaných entít pre slovenský jazyk. In: Znalosti 2012. Sborník příspěvků 11. ročníku konference. Praha: Matfyzpress 2012, s. 52 – 61.

STRAKOVÁ, Jana – STRAKA, Milan – HAJIČ, Jan: Open-source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore, Maryland: Association for Computational Linguistics 2014, s. 13 – 18.

STRAKOVÁ, Jana – STRAKA, Milan – ŠEVČÍKOVÁ, Magda – ŽABOKRTSKÝ, Zdeněk: Czech Named Entity Corpus. In: Handbook of Linguistic Annotation. Dordrecht: Springer 2017, s. 855 – 873.

ŠEVČÍKOVÁ, Magda – ŽABOKRTSKÝ, Zdeněk – KRŮZA, Oldřich: Named Entities in Czech: Annotating Data and Developing NE Tagger. In: Text, Speech and Dialogue 2007. Eds. V. Matoušek – P. Mautner. Heidelberg: Springer 2007, s. 188 – 195.

Slovenský národný korpus – r-mak-6.0. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2017. Dostupný na: https://korpus.sk/ver_r(2d)mak.html

[9] https://www.european-language-grid.eu/