Szavak mondatkörnyezet- és gráf alapú beágyazásainak összehasonlítása és kombinálása

Kardos Péter
III. évf. programtervező informatikus

Témavezető: Dr. Farkas Richárd
SZTE TTIK Számítógépes Algoritmusok és Mesterséges Intelligencia Tanszék

A természetes nyelv feldolgozás (NLP) egyik alapkövévé vált a szavak folytonos vektorokkal való reprezentálása. Ezek az úgynevezett szó beágyazások (*word embedding*) lehetővé teszik, hogy a hasonló jelentésű szavakat közeli vektorokkal (valamilyen vektorhasonlósági metrikát használva) írjunk le. Több ilyen előtanított szó beágyazási modell létezik, melyeket szövegekből építettek fel. Manapság egyre kifinomultabbak a gráfokat folytonos térbe leképző módszerek, melyekkel akár tudásbázisokat - gráfokként felfogva - ágyazhatunk be vektorterekbe ha a gráf csúcsai szavakat reprezentálnak, a köztük lévő kapcsolatokat pedig címkézett élek.

Munkámban a szavak mondatkörnyezet- és gráf alapú beágyazását hasonlítom össze. Két tudásbázis segítségével hozom létre a gráf-alapú szó beágyazásokat: WordNet és ConceptNet. A Node2Vec és Diff2Vec algoritmusokat használom, hogy az előbbi tudásbázisokból - mint gráfokból - beágyazásokat készítsek.

Ezeket jobb eredmények érdekében lehet egymással és mondatkörnyezet alapú beágyazásokkal is kombinálni. A dolgozatban újszerű algoritmust javaslok, implementálok és értékelek ki a különböző típusú beágyazások előnyeinek kiaknázására. Az előálló beágyazásokat három szóasszociációs és egy hiperníma erősséget mérő feladatokon értékeltem ki.

A számtalan empirikus kísérlet eredményeként azt találtam, hogy a tudásbázisokból gráf alapú módszerekkel kapott beágyazásokat kombinálva a mondatkörnyezet alapú beágyazásokkal jobb eredményeket ér el mintha azokat külön használnánk.