

EÖTVÖS LORÁND TUDOMÁNYEGYETEM

INFORMATIKAI KAR

DIPLOMAMUNKA TÉMABEJELENTŐ

Hallgató adatai:

Név: Kántor Attila

Neptun kód: EGS78S

Képzési adatok:

Szak: programtervező informatikus, mesterképzés (MA/MSc)

Tagozat: Nappali

Belső témavezetővel rendelkezem

Témavezető neve: Grad-Gyenge László

munkahelyének neve: Black Sheep Intelligence Kft.

munkahelyének címe: 1088, Budapest Rákóczi út 10. 3/1a

beosztás és iskolai végzettsége: Programtervező-matematikus

A diplomamunka címe: Szemantikus reprezentáció magyar nyelv esetén

A diplomamunka témája:

(A témavezetővel konzultálva adja meg 1/2 - 1 oldal terjedelemben diplomamunka témájának leírását)

A természetes nyelvfeldolgozás (NLP) a mesterséges intelligencia azon ága, amely az emberi nyelv feldolgozásával foglalkozik. Az NLP egy népszerű területe a szemantikus reprezentáció. A szemantikus reprezentációs algoritmusok természetes nyelven írott szöveg valamely formáját képezik le egy vektortérbe, annak jelentéstartalma alapján. Ezen vektorok később különböző módon felhasználhatók, például chatbot-ok implementálására, szövegbányászatra, információ visszakeresésre. A reprezentációs algoritmusok az akadémia és az ipar által is előszeretettel használt eszközök. A mesterséges intelligencia fejlesztésben élen járó vállalatok által prezentált módszerek kiemelkedően teljesítenek néhány NLP feladaton. Ilyen algoritmus például a Facebook által publikált InferSent, vagy a Google által kiadott USE, vagy BERT. A szemantikus reprezentációs módszerek többsége nyelvfüggő, azaz csak olyan nyelven íródott természetes szövegekre lehet őket alkalmazni, amely nyelven tanították őket. A tanításhoz jelentős mennyiségű szöveg szükséges. A tapasztalat azt mutatja, hogy a módszerek pontosabb eredményt érnek el, ha a tanítás címkézett adatokon (supervised leaning) történik. Míg a szélesebb körben beszélt nyelveken többnyire elérhető nagy mennyiségű címkézett tanítóadat, addig a kevésbé ismert (pl.: magyar) nyelveken a hiányuk akadályokat állít fel. Diplomamunkám célja a meglévő módszerek vizsgálata, majd olyan szemantikus reprezentációs algoritmusok fejlesztése, amelyek áthidalva ezen akadályokat képesek a kevésbé populáris nyelvek szemantikáját ábrázolni. Továbbá ezen módszerek kiértékelése és összehasonlításos vizsgálata.

Budapest, 2019.12.01.