Novák Attila (PPKE ITK)
Középmagyar disztribúciós szemantika

Az utóbbi évtizedben a digitális formában rendelkezésre álló adatok mennyiségének és a számítógépek számítási teljesítményének növekedése olyan paradigmaváltást hozott a számítástudományban, amely a nyelvtechnológiát is új alapokra helyezte. Az évtizedekkel korábban kidolgozott, de hosszú időn keresztül csak játékszernek tekintett neurális modellek komplexitása és teljesítménye ugrásszerűen nőtt ezen fejlemények hatására, és mára elsöpörték a korábbi szabályalapú és statisztikai modelleket. A nyelv modellezése terén empirikus megerősítést kapott a strukturalisták elképzelése arról, hogy a nyelvi elemek disztribúciója szolgál a legfontosabb információforrásként a grammatikai tudás számára.

Az előadásban egy egyszerű (és mára bizonyos értelemben "meghaladott") neurális modellt mutatok be, amely egyszerűsége ellenére meglepő mélységben és minőségben meg tudja ragadni a szavak szemantikai, szintaktikai és morfológiai jellegzetességeit egy sokdimenziós vektortérbeli reprezentációt hozva létre. A szemléltetésre fogunk koncentrálni: a modell kétdimenziós vetületében kalandozva viszonylag jól demonstrálható, hogy nem zagyvaság, ami létrejött. Ez a mi esetünkben talán azért meglepőbb egy kicsit a szokásosnál, mert viszonylag kevés adatból kellett dolgozni: az 1 millió szavas, élőnyelvihez közeli ó- és középmagyar levél- és perszövegekből épített Történeti Magánéleti Korpusz (TMK) szóanyaga alkotja modell alapját. Ennyi adat ennek a modellnek általában nem elég ahhoz, hogy meggyőző eredményt kapjunk, ezért szükség volt bizonyos trükkökre, amelyekre az előadásban kitérek. Az eredmény pedig egyfajta "mentális" jellegű térkép a korpusz lexikai anyagához, amelyből közvetlenül kezdeményezhető az egyes lexémák kontextuális előfordulásait visszaadó korpuszlekérdezés.