

A szerkezeti egységek kontextusmeghatározó szerepe panaszlevelek gépi feldolgozásában

Kutatásunk célja egy olyan rendszer kidolgozása, amely az ember-gép kommunikációt egy korlátozott szinten végrehajtja, ezzel megvalósítva a géppel támogatott megértést. Egy olyan automatikus kategorizáló rendszer kidolgozása a cél, amely egy adott intézmény ügyfélszolgálatára beérkező panaszleveleket készít elő a humán feldolgozás számára. Az előkészítés során a rendszer nyelvi jellemzők figyelembevételével a későbbi feldolgozást elősegítőpanasz kategóriákba sorolja az egyes leveleket, majd egyenesen az adott kategóriában jártas ügyintézőhöz továbbítja.

A feladat sajátossága, amely egyben a kihívást is jelenti, hogy a levelek nagy mennyiségben tartalmaznak olyan információt, amelynek sem a kategorizálás, sem pedig a további érdemi ügyintézés szempontjából nincs jelentősége. A levelek hétköznapi nyelven íródtak, megfogalmazásuk sokszor pontatlan, továbbá az egyes közigazgatási kategóriákra jellemző terminusok előfordulása rendszertelen, használatuk pedig sokszor nem felel meg a szaknyelvi használatnak. A levelekben előforduló irreleváns információk túl tág kontextust idéznek fel, amelyet nehéz kezelni. Ennek a problémának a megoldására vélem alkalmasnak a levelek szerkezeti egységekre való bontását, úgy mint Bemutatkozás, Probléma, Egyéb körülmények, Lezárás stb. A szerkezeti egységek szűkebb kontextusként szolgálnak a bennük előforduló információ értelmezéséhez, megmutatják, hogy milyen jellegű információkat a levél mely részében érdemes keresni, valamint lehetővé teszik a kategorizálás, valamint a további ügyintézés szempontjából irreleváns részek figyelmen kívül hagyását. A szerkezeti egységek továbbá egyszerre top-down és bottom-up információkinyerést tesznek lehetővé azáltal, hogy kontextusként szolgálnak a bennük előforduló lexikális információk interpretálásához, amelynek elméleti jelentősége, hogy a rendszer a kontextuális és a nyelvi információkat együttesen veszi figyelembe.

A feladat egy kulcsalapú modell (cue-based model) használatával valósul meg, amely a panaszlevelekre jellemző mikrogrammatikára épül. A használt kulcsalapú modell egy olyan valószínűségi modell (probabilistic model), amely egyrészt azokat a kulcskifejezéseket (n-grammokat) keresi, amelyek statisztikailag jellemzik az egyes szerkezeti egységeket és panasz kategóriákat, másrészt a különböző forrásokból származó kulcsokat (együtt előfordulási alapon) egy valószínűségi értékkel kombinálja, majd ez alapján hajtja végre a kategorizálási feladatot (Jurafsky 2004). A kategorizálási feladatot egy olyan forgatókönyv koordinálja, amely azokat a lépéseket tartalmazza, amelyek szükségesek ahhoz, hogy a kategorizálás szempontjából releváns szerkezeti egységek, illetve az ezekben megfigyelhető mikrogrammatika alapján a kategorizálás megtörténjen.

Az előadásban azt szeretném bemutatni, hogy egyrészt a levelekben megfigyelhető hétköznapi nyelvhasználat okozta túl tág kontextus szűkítéséhez milyen mértékben járulhatnak hozzá a szerkezeti egységek, másrészt, hogy ez alapján miként valósítható meg a kategorizálási feladat.

Hivatkozás:

Jurafsky, D. 2004. Pragmatics and Computational Linguistics. In: Horn, L.R. - G. Ward (szerk.) *The Handbook of Pragmatics*. Cambridge: MA, MIT Press. 578–604.