Szövegszinkronizációs módszerek, hibrid bekezdés- és mondatszinkronizációs megoldás (kivonat)

Pohl Gábor
Pázmány Péter Katolikus Egyetem Információs Technológiai Kar
pohl@morphologic.hu

Szövegszinkronizáción (text alignment) két- vagy többnyelvű szövegekben az egymás fordításának tekinthető szövegegységek meghatározását értjük. Fordítómemória-alkalmazásokhoz, fordítások terminológiai konzisztenciájának ellenőrzéséhez, párhuzamos (szinkronizált) korpuszok építéséhez legalább mondatszintű szinkronizációra van szükség. Az egymás fordításának tekinthető mondatok meghatározása nehéz feladat, mivel a fordítók a mondatok határait a fordítás során megváltoztathatják, (véletlenül) elhagyhatnak, illetve beszúrhatnak mondatokat.

Egy konkrét szinkronizációs megoldás ismertetésén túl az előadás célja betekintést adni a szövegszinkronizáció témakörébe. Először bemutatjuk és értékeljük a legfontosabb eddig publikált szinkronizációs módszereket, majd az eddigieknél szakszövegek esetében pontosabb, hibrid bekezdés- és mondatszinkronizációs megoldást ismertetünk.

Az általunk kidolgozott hibrid megoldás során egy alkalmas heurisztika használatával kombináljuk a mondatok hosszainak arányát és a szövegpárban talált, egymásnak megfeleltethető pontokat (horgonyokat) kihasználó módszereket. A mondathosszakon alapuló módszer horgonyok hiányában is robosztussá teszi a hibrid megoldást, a kritikus részek (például beszúrások) környezetében viszont az esetlegesen előforduló horgonyok tehetik pontosabbá a szinkronizációt.

Azért, hogy magyar nyelvű szövegek esetén jól alkalmazható legyen a módszer, a horgonyjelölteket szótövesítve is keressük. A horgonyjelöltek közül a hibrid algoritmus által használt horgonyokat – a kellő pontosság érdekében – Ribiero és társai által már korábban alkalmazott statisztikai módszerekkel választjuk ki.

Kulcsszavak:

szövegszinkronizáció (text alignment), mondatszinkronizáció (sentence alignment), bekezdésszinkronizáció (paragraph alignment) horgony, statisztikai szűrés, dinamikus programozás