Tóth Krisztina
Közgazdasági programozó matematikus, V. évf.
Mesterséges Intelligencia Kutatócsoport

Konzulens: dr. Kocsor András
Tudományos főmunkatárs

HIBRID ALGORITMUS MAGYAR-ANGOL NYELVŰ SZÖVEGÁLLOMÁNYOK MONDATSZINTŰ MEGFELELTETÉSÉRE

Az MTA Mesterséges Intelligencia Kutatócsoportjánál magyar-angol fordítóprogram fejlesztéséhez magyar-angol párhuzamos korpusz építését és feldolgozását kezdtük meg. A feldolgozás első lépéseként a mondatok párhuzamosítását oldottuk meg. Ez a dolgozat egy hatékony hibrid párhuzamosítási technikát mutat be a szövegek párhuzamosítására.

A szövegszinkronizáció két típusát különböztetjük meg a szövegek lefedettsége alapján: a teljes és részleges szinkronizációt. Az előbbi többnyire a szövegegységek valamilyen mérték szerinti hosszán alapul. A részleges szinkronizáció során pedig olyan lexikai információkat (horgonyokat) keresünk a szövegekben, amelyek mind a magyar, mind az angol nyelvben megtalálhatók. A fent említett két szinkronizációs módszert ötvözve egy olyan hibrid megoldáshoz jutunk, amelyben a statisztikai információkat tartalmazó mondathossz-alapú összerendelés kiegészül a részleges szöveg-összerendelés alapjául szolgáló horgonykeresési eljárással.

Az általunk elkészített algoritmus megvalósítása során azt az általános fordítói tapasztalatot vettük alapul, hogy a mondathatárok nem nyúlhatnak át a bekezdéshatárokon. Hosszmértéknek a karakterszámot választottuk, kihasználva, hogy a mondatok karakterszáma korrelál, figyelembe véve, hogy a magyar szövegek átlagosan 15 százalékkal hosszabbak, mint az angol megfelelőik. Természetesen csak legegyszerűbb esetben fordul elő, hogy egyetlen forrásnyelvi mondatnak egy célnyelvi mondat felel meg. Előfordulhatnak - a fordítói szabadságnak köszönhetően – 1-N, N-1 és N-M megfeleltetések is; ezek felismerésére az elkészült algoritmus alkalmas. A magyar nyelvre publikált horgonykereső eljárások a számok normalizált alakját és a nagykezdőbetűs szavakat használják horgonynak, míg az általunk bemutatott eljárás kezeli a mozaikszavakat, rövidítéseket és a nagybetűs szavak helyett a tulajdonneveket. A tulajdonneveket a Mesterséges Intelligencia Kutatócsoport által készített kvázi-nyelvfüggetlen tulajdonnév felismerővel nyertük ki.

Egy mondatszinkronizáló algoritmus eredményessége jelentősen függ az összerendelést megelőző mondatszegmentálástól. Ezért kidolgoztunk egy mondatszegmentáló algoritmust, amely kevesebb, mint 0.5%-os hibával dolgozik mind angol, mind magyar nyelvű szövegeken. Az elkészített mondatszegmentáló algoritmus a Szeged Treebank véletlenszerűen kiválasztott részkorpuszain 99.7-99.85%–os eredményt ért el. Ez az eredmény meghaladja a szakirodalomban publikált, magyar nyelvre készített eredményeket.

A hibrid algoritmus a horgonyokat tartalmazó mondatokban nagyobb pontossággal határozza meg a szinkronizációs egységeket, míg horgonyok hiányában a hosszúságalapú összerendelést követi. Ezáltal pontosabb eredményt értünk el, mint a referenciaként használt Gale és Church hosszúságalapú összerendelés.