Doménspecifikus korpusz építése és validálása

Bevezetés

Napjainkban az interneten fellelhető tartalmak tömegesen elérhetők, és némi ráfordítással milliárdos nagyságrendű korpuszok állíthatók elő. Ezek előnyei, hogy sokféle nyelvi jelenség megfigyelhető bennük, hátrányuk viszont, hogy sokféle szöveget tartalmaznak, így nem tudunk például ezeken tanítani és építeni olyan ú. n. doménspecifikus osztályozókat, amelyek csak az adott típusú szövegre jellemzőek.

A terminuskivonatoló alkalmazás fejlesztésének első lépése egy az adott domént jellemző korpusz készítése, lehetőleg minél relevánsabb és minél több anyagból (Nagy 2012). Elegendő nagyságú korpusz (pl. a szóvektoralapú elemzésekhez [Grefenstette–Muchemi 2016]) kézzel történő gyűjtése azonban rendkívül időigényes.

Jelen tanulmány a disszertációmhoz szükséges korpusz építésének módszertani vizsgálata, amelyben az ökoinnováció doménhez tartozó szövegeket gyűjtök és értékelek, oly módon, hogy hasonlóságot mérek a letöltött anyagok között. Ezzel célom egy nagy méretű, ugyanahhoz a doménhez tartozó korpusz készítése a későbbi munkához.

Hipotéziseim a következők: 1) A letöltött szövegek elérik a 0,50-es hasonlósági értéket. 2) Minél kevesebb szövegszóból áll a szöveg, annál kevesebb lesz a hasonlósági értéke a referencia korpuszhoz képest.

Módszertan

A szövegek letöltése egy filtereket alkalmazó programkóddal történt, amelyet angol nyelvre Greffenstette és Muchemi (2016) írt meg. Ennek magyarra történő átültetése után szükség volt egy kezdő URL-lista, illetve egy mintákat tartalmazó fájl összeállítására. A kezdő, manuálisan összeállított lista alapján a program letöltötte a szövegeket, amelyeket szűrt az előre megadott minták alapján. A validáláshoz 10%, tehát 20 darab random szöveget választottam, azzal a kikötéssel, hogy a szövegnek több, mint 500 szövegszóból kell állnia. Referencia szövegeknek 4 szöveg lett kijelölve. A szövegekből a lemmatizálás után kiszűrtem a stopszavakat, majd az XLike projekt kereteiben készült összehasonlító programmal (Rettinger et al. 2012) számoltam hasonlóságot és statisztikailag elemeztem az adatokat.

Eredmény

A kiválasztott szöveg átlag hossza stopszavak nélkül 1428 szövegszó, szórásuk 1054,49. A referenciaszövegek hossza: 2960, 4169, 2483, 10133 szövegszó. A hasonlóság mérésekor kapott értékek 0 és 1 közé esnek.

Az eredmények referencia korpuszonként a következő átlagot hozták: 1: 0,524; 2: 0,373; 3: 0,541; 4: 0,297. A szórásuk 1: 0,198; 2: 0,192; 3: 0,201; 4: 0,14.

A referencia szövegek hasonlóságát is mértem egymáshoz viszonyítva. Az 1-es és a 4-es, a 3-as és a 4-es és a 2-es a 3-as szövegek szignifikánsan eltérnek, míg az 1-es és a 2-es között tendencia van.

Az első hipotézisem nem teljesült, mivel egyik referencia korpuszhoz mért átlag hasonlósági érték sem lett magasabb, mint 0,50.

A második hipotézisemnél a szövegek mérete és az eredmény között van egy gyenge negatív kapcsolat (-0,4, p=0,07).

Következtetések

Bár az általam használt szkripten kívül más hasonló célú alkalmazások is léteznek (pl. Remus–Biemann 2016), ennek előnye, hogy egyszerűen átültethető más nyelvekre, illetve előzetesen nincs szükség a szövegek elemzésére, csupán egy jól átgondolt mintákat tartalmazó fájl készítésére. A validálásra a jól megválasztott minták használata mellett is szükség lehet.

Az eredmények azt mutatták, hogy a szövegek hasonlósági értékei nem voltak átlagosan 0.50-nél magasabbak, ennek hátterében az állhat, hogy bár a szöveg tartalma részben lefedi a domént, de vagy túl specifikus vagy túl általános, illetve megjegyzendő, hogy az ökoinnováció számos aldoménnel rendelkezik, különböző területeket érint (ez elmondható a referencia korpuszok közötti eltéréssel kapcsolatban is). Ennek mérésére egy későbbi, kiterjesztett kutatás szolgálhat, ahol emberi annotátorok döntik el, hogy a szöveg milyen mértékben tartozik a doménhez, illetve miért.
Az eredmények szerint továbbá látszik, hogy a szövegek hossza kis mértékben, de befolyásolja a hasonlóságot, ennek további vizsgálata a minta növelését igényli.

Hivatkozott irodalom

Grefenstette, G. – Muchemi, L. 2016. Determining the Characteristic Vocabulary for a Specialized Dictionary using Word2vec and a Directed Crawler. *Lexicographic Resources for Human Language Technology GLOBALEX 2016 Workshop Proceedings,* 24 May 2016, Portoroz. https://arxiv.org/pdf/1605.09564v1.pdf Letöltés: 2016. 12. 03.

Nagy Á. 2012. *Terminológiakivonatolás francia nyelvű szabadalmi leírásokból szabály alapú és statisztikai módszerek segítségével.* PhD-értekezés. Szegedi Tudományegyetem Bölcsészettudományi Kar, Nyelvtudományi Doktori Iskola. doktori.bibl.u-szeged.hu/1768/1/disszertacio_NagyAgoston_egyben.pdf Letöltés: 2016. 05. 15.

Remus, S. – Biemann, Ch. 2016. Domain-Specific Corpus Expansion with Focused Webcrawling. *Proceedings of the 10th edition of the Lexicographic Resources for Human Language Technology 23-28 May 2016, Portorož (Slovenia).* http://www.lrec-conf.org/proceedings/lrec2016/pdf/316_Paper.pdf Letöltés: 2016. 12. 03.

Rettinger, A. – Zhang, L. – Rupnik, J. – Muhič, A. 2012. *Cross-lingual document linking prototype.* Deliverable D4.1.1. XLike project. http://cordis.europa.eu/docs/projects/cnect/2/288342/080/deliverables/001-D411Crosslingualdocumentlinkingprototype.pdf Letötés: 2016. 12. 01.