Bizonytalanságot jelölő kifejezések és hatókörük azonosítása természetes nyelvi szövegekben: a CoNLL-2010 verseny tapasztalatai

Farkas Richárd[1,2], Vincze Veronika[2], Móra György[2],
Csirik János[1,2], Szarvas György[1]

[1] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
{rfarkas, csirik, szarvas}@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
{vinczev, gymora}@inf.u-szeged.hu

Kivonat: A CoNLL-2010 konferenciához kapcsolódó nemzetközi versenyfeladat a bizonytalanságot jelölő kifejezések, és azok hatókörének azonosítását tűzte ki célul angol nyelvű szövegekben. Cikkünkben bemutatjuk a versenykiírást, a beérkezett rendszereket, a kiértékeléshez épített adatbázisokat, értékeljük az eredményeket, végül pedig beszámolunk egy – hasonló elvek alapján épített – magyar nyelvű, bizonytalanságot jelölő kifejezésekre annotált korpuszról.

1 Bevezetés

A CoNLL-2010 konferenciához kapcsolódó nemzetközi versenyfeladat a Szegedi Tudományegyetem Informatikai Tanszékcsoportjának szervezésében a bizonytalanságot jelölő kifejezések, és azok hatókörének azonosítását tűzte ki célul angol nyelvű szövegekben [1]. A feladat fontossága abban rejlik, hogy a különféle számítógépes nyelvészeti alkalmazásokban lényegi szerep jut a tényszerű és a bizonytalan, illetve tagadott információ megkülönböztetésének, hiszen például információkinyerés és szemantikus keresés esetében a felhasználónak többnyire tényszerű információra van szüksége, így alkalmazástól függően a rendszer vagy kiszűri a bizonytalan / tagadott szövegrészeket, vagy pedig a tényektől elkülönítve adja őket vissza a felhasználónak.

Cikkünkben összefoglaljuk a verseny tapasztalatait, valamint beszámolunk egy magyar nyelvű, bizonytalanságot jelölő kifejezésekre annotált korpuszról.

2 A versenyfeladatok

A bizonytalanságot tartalmazó szövegrészek azonosítása történhet mondatszinten és hatókör szinten. Az első esetben elégséges a mondatról eldönteni, hogy az tartalmaz-e bizonytalan információt vagy sem, míg a második esetben a cél: megjelölni a mondaton belül a bizonytalanságot jelző nyelvi elemeket (kulcsszavakat) és azok mondatbeli

hatókörét. Noha az utóbbi feladat nagyobb kihívást jelent, a legtöbb alkalmazás számára mégis előnyt jelent ez a jelölési módszer, hiszen lehetnek olyan (általában összetett) mondatok egy szövegben, ahol a mondat egy része bizonytalan információt hordoz, más részében viszont hasznos tényszerű információ rejlik.

A versenykiírásban szereplő két feladat a fentieknek megfelelően mondat- és hatókör szintű címkézést tűzött ki célul. Az első feladat mondatszintű címkézést kívánt meg aszerint, hogy a mondat tartalmaz-e bizonytalan információt vagy sem. A rendszereknek biológiai témájú cikkek, illetve Wikipédia-szócikkek mondatait kellett osztályozniuk. A második feladatban biológiai cikkekben kellett bejelölni a kulcsszavakat és azok mondaton belüli hatókörét.

A versenyfeladatokhoz biztosítanunk kellett tanító és kiértékelő adatbázist is. Tanító adatbázisként a biológiai doménre a BioScope korpusznak [2] a tudományos cikkek absztraktjait és teljes cikkeket tartalmazó részét választottuk, a kiértékeléshez pedig újonnan annotáltunk 15 biológiai témájú cikket. A Wikipédia doménen pedig mind a tanító, mind a kiértékelő adatbázist az angol nyelvű Wikipedia *weasel* címkével ellátott (homályos, kétértelmű, túlzó vagy félrevezető információt tartalmazó) bekezdései közül válogattuk ki, melyekben kézzel megjelöltük a bizonytalanságot jelző szavakat. Néhány példamondat az adatbázisokból (<> jelöli a kulcsszavakat, míg () a hatóköröket):

The album, which was recorded in less than two weeks, contains <arguably> the band's two <most famous> songs, "Wonderwall" and "Don't Look Back in Anger" and their first UK #1 single, "Some Might Say". (Wikipédia)

Thus, misregulation of these genetic pathways (<may> confer unrestricted proliferative capacities to a range of glial cell types), but (how this occurs remains <unclear>). (biológiai publikáció)

Megjegyezzük, hogy a wikipédiás példamondat egyben azt is mutatja, hogy a kulcsszójelöltek nem minden előfordulásukban szerepeltek ténylegesen kulcsszóként: a *some* és a *might* gyakori kulcsszavak, de a fenti példában egy dal címének részeként – azaz metanyelvi használatban – nem utalnak bizonytalanságra.

A versenyzőknek lehetőségük nyílt az általunk rendelkezésekre bocsátott adatbázisok mellett további erőforrások használatára is a rendszerük fejlesztése során.

3 Versenyeredmények, értékelés

A versenyen 23 intézet kutatói vettek részt a világ minden tájáról. Az első feladat biológiai részére 22 csoport, a wikipédiás szövegek feldolgozására 16 csoport, míg a második feladatra összesen 13 csapat vállalkozott.

Az első feladat kiértékelése mondatszinten történt: a bizonytalan osztály F-mértékét alkalmaztuk mint fő kiértékelési metrikát. A második feladatban, ahol a kulcsszavakat és azok hatókörét is azonosítani kellett, egy szigorú, hatókör szintű kiértékelési metrikát használtunk: pontos találatnak csak azt fogadtuk el, ahol a kulcskifejezések és hatóköreik is pontosan lettek megállapítva.

A legjobb rendszerek az első feladatban 86% (biológiai domén), illetve 60% (Wikipédia) körüli F-mértéket értek el, a másodikban pedig 57% körülit. Utóbbi eredmény egyrészt a feladat nehézségét, másrészt pedig a kiértékelési metrika szigorúságát is jelzi: bizonyos esetekben a hatókörök rugalmasabb illeszkedése lenne kívánatos (például írásjelek, hivatkozások, zárójeles megjegyzések kezelése).

Az első feladatra a legjobb eredményt elért versenyzők biológiai szövegeken szekvenciajelöléses megközelítést alkalmaztak, míg a Wikipédia-szövegeken a szózsák típusú modellek bizonyultak sikeresnek.

4 Bizonytalanságot jelölő kifejezések a magyarban

A verseny résztvevőinek magas száma arra utal, hogy a bizonytalan szövegrészek azonosításának problémája élénken foglalkoztatja a számítógépes nyelvész kutatókat világszerte. Míg az eddigi kutatások nagy része az angol nyelvre irányult (azon belül is elsődlegesen az orvosi-biológiai szövegekre), szeretnénk a továbbiakban a magyar nyelvre is kiterjeszteni az ilyen témájú kutatásokat. E cél érdekében kísérleti jelleggel építettünk egy magyar nyelvű, Wikipédia-szócikkekből álló adatbázist[1], melyben kézzel annotáltuk a bizonytalanságot jelölő nyelvi elemeket, az ún. weasel szavakat[2]. A weasel szavak a véleményeket megfelelő forrás vagy alátámasztás nélkül tálalják: nem tükrözik egy enciklopédia szerkesztői (és olvasói) által elvárt semleges stílust. A következő példában az információ forrása nincs megadva, pusztán a *sokan* kifejezés utal a vélemény hordozójára:

*Ma már **sokan** úgy vélik, hogy ez a megítélés erősen szubjektív, hiszen Linné maga is rendszerint a svédországi fajt (vagy alfajt) látta a „legtipikusabbnak”.*

A korpusz létrehozásában követtük az angol nyelvű adatbázis építésekor alkalmazott alapelveket annak érdekében, hogy az eredményeket összevethessük a versenyfeladathoz épített korpusz adataival. Elsőként egy nyelvészeti szempontok alapján összeállított kulcsszólista segítségével gyűjtöttünk a magyar Wikipédia szócikkeiből bekezdéseket, majd ezekből – a kulcsszó-jelöltek gyakorisági adatait szem előtt tartva – véletlenszerűen válogattuk ki az annotálandó bekezdéseket. A végső annotált korpusz 1710 bekezdést tartalmaz. A munka során a nyelvészek bejelölték a weasel kifejezéseket a szövegekben, majd azokat a mondatokat minősítettük bizonytalannak, amelyek legalább egy kulcsszót tartalmaznak. A 11647 mondatból 953 volt ilyen (8,18%). Összehasonlításképpen: az angol tanító adatbázison 22,36%, a kiértékelő adatbázison pedig 23,19% volt a bizonytalan mondatok aránya.

A szövegekben összesen 1156 kulcsszó fordult elő, vagyis egy bizonytalan mondat átlagosan 1,21 kulcsszót tartalmazott. A leggyakoribb kulcsszavak, illetve kulcskifejezések a következők voltak: *számos N* (132 előfordulás), *valószínűleg* (128), *egyes N* (91). A kulcsszavak csoportjait tekintve elsődlegesen a határozatlan vagy általános

[1] A korpusz Creative Commons licenc alatt elérhető a www.inf.u-szeged.hu/rgai/uncertainty oldalon.

[2] http://hu.wikipedia.org/wiki/WP:WEASEL

kvantorokat (*egyes, néhány*) tartalmazó kifejezések és a bizonytalanságra vagy általánosságra utaló határozószók (*valószínűleg, feltehetőleg, általában*) domináltak a korpuszban, de a *más N* és *mások* kifejezések használata is jellemző volt.

A fenti számadatok azt mutatják, hogy a magyar Wikipédiában a bizonytalan mondatok aránya az angollal összevetve jelentősen kisebb. Ennek két fő oka lehet. Egyrészt, az angol Wikipedia szerkesztői közössége valószínűleg sokkal heterogénebb, mint a magyaré, ezért ott nagyobb az esély arra, hogy egy új szócikket egy kevesebb szerkesztői tapasztalattal rendelkező tag hozzon létre, növelve ezzel a bizonytalan szócikkek arányát. Másrészt, a Wikipédiák méretbeli különbségéből adódóan a bizonytalan szócikkek száma abszolút értékben véve jóval kevesebb a magyarban, azaz a szerkesztők könnyebben és gyorsabban ki tudják ezeket javítani.

A kulcsszavak gyakoriságát illetően a két nyelv között nincs számottevő eltérés. Az angol adatbázis leggyakoribb kulcsszavai a *some*, *may* és *others* voltak, míg a magyarban is gyakran fordultak elő a *számos*, *egyes*, *más*, *mások* kifejezések. Mivel a magyar nyelv morfológiailag fejezi ki a ható modalitást, az angol pedig a *may* segédigével, a *may* kulcsszó gyakorisága a *–hAt* morfémát tartalmazó elemek gyakoriságával vethető össze, ez pedig a két nyelv esetében hozzávetőlegesen megegyezik.

5 Összegzés

Tanulmányunkban beszámoltunk a CoNLL-2010 konferenciához kapcsolódó versenykiírásról, ahol a cél bizonytalanságot jelölő kifejezések azonosítása volt. Röviden bemutattuk a kiértékeléshez épített adatbázisokat, ismertettük a beérkezett rendszereket, végül pedig leírást adtunk egy – hasonló elvek alapján épített – magyar nyelvű, bizonytalanságot jelölő kifejezésekre annotált korpuszról, mely a későbbiekben a magyar nyelvre fejlesztendő, bizonytalan szövegrészeket azonosító alkalmazások tanításában, illetve egységes kiértékelésében tölthet be fontos szerepet.

Köszönetnyilvánítás

A kutatást – részben – a TEXTREND, a BELAMI és a MASZEKER kódnevű projektek keretében az NKTH támogatta.

Bibliográfia

1. Farkas, R., Vincze, V., Móra, Gy., Csirik, J., Szarvas, Gy.: The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, Uppsala (2010) 1–12
2. Vincze, V., Szarvas, Gy., Farkas, R., Móra, Gy., Csirik, J.: The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes. Bioinformatics Vol. 9, No. 11 (2008)