Gyógyszermellékhatások kinyerése magyar nyelvű orvosi szaklapok szövegeiből

Farkas Richárd[1], Miklós István[1], Tímár György[2], Zsibrita János[1]

[1] Szegedi Tudományegyetem, Informatikai tanszékcsoport
Szeged, Árpád tér 2.
{rfarkas,mikist,zsibrita}@inf.u-szeged.hu
[2] Comfit Kft.
gyorgy.timar@comfit.hu

1 Kivonat

A magyar nyelvű orvosi szaklapok gyakran közölnek ún. esettanulmányokat, melyben leírják, hogy bizonyos hatóanyagok (ágensek), például gyógyszerek hatására pácienseknek milyen tünetei (szimptómák) voltak megfigyelhetőek. A Comfit Kft. és a Szegedi Tudományegyetem együttműködésében megvalósuló projekt azt tűzte ki célul, hogy megvizsgálja a nyelvtechnológia lehetőségeit és a jelenleg kézi átolvasással történő elemzése helyett egy automatikus megoldást szolgáltasson.

A projekthez rendelkezésre állt 4600 magyar nyelvű orvosi szakcikk és a bennük megtalálható ágens-tünet összerendelések, melyet a Comfit Kft. munkatársai az elmúlt években manuálisan gyűjtöttek.

Első lépésben a PDF formátumban lévő újságcikkekből kellett a szöveges tartalmakat kinyerni. Itt komoly gondot okozott a többhasábos szerkesztés és a grafikonok, hirdetések nagy száma. A hasábok azonosítására egy, a szóközök sűrűségét figyelő algoritmussal találtunk megoldást. Egy következő lépésben a dokumentumokat megszűrjük. Például kidolgoztunk egy gépi tanulási módszereken alapuló bibliográfiablokk-azonosító modult, amire azért volt szükség, mert a hivatkozások címei gyakran tartalmaztak ágensmegnevezéseket, ami félrevezette az információkinyerő rendszert.

Maga az információkinyerő rendszer három nagy részből áll össze. Először a szövegben azonosítjuk a potenciális hatóanyag- és tünetemlítéseket, majd megállapítjuk, hogy melyik tünet mely hatóanyagokra vonatkozik és végül az azonosított említéseket az adatbázis azonosítóira kell leképeznünk.

Az egyik legnagyobb problémát az okozta, hogy a tanító adatbázis egy szövegbeli előfordulástól független adattáblaként állt rendelkezésre, míg az információkinyerő rendszernek szövegbeli említések pontos helyének (mondatkörnyezet stb.) ismerete szükséges. Hogy ezeket az említéseket azonosítsuk szótárakat és rövidítéslistákat illesztettünk a szövegekre [1]. Ezeket használtuk utána a gépi tanuló algoritmusok tanító adatbázisaként. A tanulás során keletkezett szekvenciajelölő modell használatával még nem látott újságcikkekben is felismerhetjük a hatóanyagokat és szimptómákat [2]. A hatóanyagok és tünetek közti relációk azonosításánál azzal a problémával szembesültünk, hogy ismét csak azonosító alapján voltak a párok jelölve, így egy párt a szövegben a két fél minden előfordulása reprezentálta. Ezért nem hagyatkozhattunk

csupán a párok egymástól független osztályozására, hanem figyelembe kellett venni minden egyedet. Erre globális optimalizálásban használt és különböző gépi tanulással segített eljárásokkal kísérleteztünk [3].

Hivatkozások

1. Farkas, R., Dobó, A., Kurai, Z., Miklós, I., Miszori, A., Nagy, Á., Vincze, V., Zsibrita, J.: Információkinyerés magyar nyelvű önéletrajzokból a nexum Karrierportálhoz. In: X. Magyar Számítógépes Nyelvészeti Konferencia (2014)
2. Móra, Gy., Farkas, R.: Szótáralapú, névelem-felismerés szóhatárainak javítása gépi tanulási módszerrel. In: VII. Magyar Számítógépes Nyelvészeti Konferencia (2010) 317–324
3. Björne, J., Ginter, F., Salakoski, T.: University of Turku in the BioNLP'11 Shared Task. BMC BioInformatics Volume 13 Supplement 11 (2011)