Magyar nyelvű hasonló tartalmú orvosi leletek azonosítása

Wieszner Vilmos[1], Farkas Richárd[1], Csizmadia Sándor[2], Palkó András[2]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2., e-mail:{wieszner,rfarkas}@inf.u-szeged.hu
[2]Szegedi Tudományegyetem, Radiológiai Klinika
6725 Szeged, Semmelweis u. 6

A radiológiai praxist támogató számítógépes nyelvészeti alkalmazás lehet egy az éppen gépelt dokumentumhoz hasonló leletek megtalálása. Demónkban egy ilyen, általunk készített rendszert mutatunk be. A kitűzött cél részszövegekre a leghasonlóbb leletek megtalálása, valamint a megtalált leletek rangsorolása.

A rangsor helyességének értékelésére rendelkezésünkre áll egy 200 elemű adatbázis, ahol az orvosok által kézzel lettek megadva a dokumentumokhoz leghasonlóbb találatok. Továbbá fontos kritérium volt, hogy a találatok között nem szerepelnek olyan leletek, melyek diagnózisa negatív.

Két lelet közötti hasonlóság kiszámítását nagyban befolyásolják az emberi tényezők, ilyenek a helyesírás, valamint a leletekből hiányzó információ vagy eltérő írásmód. A helytelen helyesírás, mint a kisbetűk, illetve gépelési hibák a figyelmetlenségből fakadnak, míg a hiányos információ és a különböző írásmód a páciens aktuális orvosán múlik. Ebből következik, hogy a rendszernek képesnek kell lennie az ilyen jellegű hibákból keletkező eltérések figyelmen kívül hagyására. A leletek tárolását a Solr rendszer [2] segítségével végeztük, ami lehetővé teszi a valós időben történő komplex kereséseket még rendkívül terjedelmes dokumentumhalmaz esetében is.

A figyelmetlenségből eredő hibák javítása könnyen megoldható [1], de a leletek közötti orvosok stílusának eltérései több kihívást rejtenek. Ha az orvos már tudja a beteg egy lehetséges diagnózisát, akkor legtöbbször nem írja le azt egy másik leletbe, valamint a leletekben eltérő lehet a rövidítések értelmezése, még ugyanazon orvos által írt leletek esetében is. A rövidítés értelmezése a kontextustól is függ, ilyen például a *CA* jelölés, ami a szövegkörnyezettől függően jelenthet szívrohamot vagy rosszindulatú daganatot a hámszöveten. A találatokat továbbá befolyásolják a tünetek, illetve a már diagnosztizált betegségek fizikai helye, valamint a mérete. A *két milliméteres csomó* az agyban, illetve a bélrendszerben teljesen más következményekkel járhat, így ezeket más méretkategóriába soroljuk, amit a hely és a nagyság határoz meg. A kór és a tünet megnevezése is változhat, leletenként, részben az orvos szóhasználatától, részben a lehetséges szinonimáktól függően. Az *infarktus* előfordulhat strokeként, vagy a bekövetkezés helyétől függően agyvérzésként is. Ennek megoldása a kontextustól függő szinonima-, illetve rövidítésfeloldás, azaz a szövegkörnyezetből kinyert részinformációkból adjuk meg a rövidítés legvalószínűbb jelentését.

Előfeldolgozó lépések után – mint például a szótövezés és írásjelek eltávolítása – a dokumentumok reprezentációját az unigramok és a kinyert numerikus

tulajdonság-érték párok alkotják. A leghasonlóbb találatokat a leleteken tf-idf normalizálással számítjuk ki, azzal a módosítással, hogy meghatározott szavak, mint a szervek megnevezése, a méretkategóriák és bizonyos előre megadott tünetek nagyobb súllyal legyenek figyelembe véve.

Köszönetnyilvánítás

Jelen kutatást a Telemedicina fókuszú kutatások orvosi, matematikai és informatikai tudományterületeken című, TÁMOP-4.2.2.A-11/1/KONV-2012-0073 számú projekt támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Hivatkozások

1. Siklósi B., Novák A., Prószéky G.: Helyesírási hibák automatikus javítása orvosi szövegekben a szövegkörnyezet figyelembevételével. In: Magyar Számítógépes Nyelvészeti Konferencia (2013)
2. Smiley, D., Pugh, E., Parisa, K., Mitchell, M.: Apache Solr 4 Enterprise Search Server (2014)