*Strokovno-znanstveni prispevek* ■

# Označevanje in odkrivanje pomenskih razmerij v medicinskih besedilih

**Špela Vintar**

**Izvleček.** Prispevek opisuje metodo samodejnega pomenskega označevanja medicinskih besedil s pojmi in pomenskimi razmerji metatezavra UMLS. Podlaga za takšno označevanje je jezikovna obdelava, pri kateri se v besedilo vnese osnovne ravni jezikoslovne analize, predvsem besedne vrste in osnovne oblike besed. V nadaljevanju opišemo dva načina izrabe tako obogatenih besedil, in sicer za namene medjezičnega iskanja dokumentov in za odkrivanje novih pomenskih razmerij v medicinskih besedilih.

# Annotating and Discovering Semantic Relations in Medical Texts

**Abstract.** The paper describes a method of automatic semantic annotation of medical texts on the basis of the UMLS Metathesaurus. Prior to semantic processing the texts are linguistically analysed, lemmatised and tagged for part-of-speech. Two application areas are then outlined in more detail, first the exploitation of semantically annotated documents in Cross-Language Information Retrieval and second the discovery of new semantic relations in medical texts.

■ **Infor Med Slov**: 2005; 10(1): 9-18

Institucija avtorja: Filozofska fakulteta, Univerza v Ljubljani.

Kontaktna oseba: Špela Vintar, Filozofska fakulteta, Univerza v Ljubljani, Aškerčeva 2, 1000 Ljubljana. e-mail: spela.vintar@ff.uni-lj.si.

## Uvod

Osnovna problematika širokih in visoko razvitih znanstvenih področij, kot je medicina, je v (ne)obvladovanju količine znanja, ki je nakopičeno v znanstvenih besedilih in člankih, pogosto zbranih v velike besedilne baze, kakršna je Medline. Računalniška obdelava jezika lahko pomembno pripomore k dostopanju do besedilnih podatkov, predvsem tako da zmanjšuje dvoumnost in variabilnost jezikovnih izrazov na vseh ravneh. Če denimo iščemo besedo *terapija*, nas navadno zanimajo tudi oblike besede *terapije*, *terapiji* itd., morda bi želeli obenem iskati tudi sopomenke, kot je *zdravljenje*, morda pa bi nas zanimala tudi besedila o iskanem pojmu v tujem jeziku, denimo angl. *treatment*, *therapy*.

Pri tradicionalnem iskanju podatkov *(Information Retrieval)* se za razreševanje oblikoslovne razvejanosti jezika uporablja krnjenje *(stemming)*, postopek, ki različne oblike iste besede "obteše" do njim skupnega jedra ali krna. Nekoliko kompleksnejši postopek jezikovne obdelave je lematizacija oziroma opremljanje besed z njihovimi pravimi osnovnimi oblikami – lemami; zanj potrebujemo leksikon besed in njihovih oblik, pri dvoumnih besednih oblikah (npr. v slovenščini *brez* kot rodilnik množine samostalnika *breza* ali predlog *brez*) pa nam pomaga besednovrstno označevanje.

Če želimo iskati dokumente v večjezičnih besedilnih zbirkah, se soočamo s problemom, da jezik iskalne zahteve ni enak jeziku, v katerem so napisani – nekateri – dokumenti. Za premostitev jezikovne vrzeli se lahko uporabi strojni prevajalnik, ki prevede bodisi le iskalno zahtevo bodisi vse dokumente v zbirki. Druga možnost, ki jo v tem prispevku podrobneje opisujemo, pa je samodejno pomensko označevanje iskalnih zahtev in dokumentov.

Nujen vir za takšno označevanje je večjezični tezaver ali ontologija, se pravi hierarhično urejena baza pojmov določenega področja, ki posameznim strokovnim terminom priredi jezikovno neodvisno pojmovno oznako. Za področje medicine je tak vir UMLS *(Unified Medical Language System)*,[a] ki v svojem metatezavru opisuje prek milijon medicinskih pojmov in vsebuje več kot pet milijonov medicinskih izrazov v različnih jezikih. Če torej pomenski označevalnik v besedilih različnih jezikov poišče medicinske izraze in jim priredi jezikovno neodvisne pojmovne oznake, in se enak postopek uporabi tudi na sami iskalni zahtevi, to omogoča medjezično iskanje dokumentov.

Pričujoči prispevek v prvem delu opisuje prototip takega medjezičnega iskalnika, ki smo ga razvili za angleški in nemški jezik v okviru projekta MUCHMORE.[b] Predstavljeni so tudi rezultati evalvacije, kjer smo pojmovno usmerjeno metodo primerjali z drugimi znanimi metodami medjezičnega iskanja. V drugem delu spregovorimo še o raziskavah, kako izrabiti pomensko označena besedila za odkrivanje novega znanja, še posebej novih pomenskih razmerij med medicinskimi pojmi.

## Pomensko označevanje za pojmovno medjezično iskanje

Temeljna naloga takšnega medjezičnega iskalnika je, da v besedilu oziroma iskalni zahtevi poišče termine in jih preslika na jezikovno neodvisno pojmovno raven. Kot vir terminologije in pojmovnih razmerij za področje medicine uporabljamo UMLS oziroma natančneje tri njegove komponente:

- **Specialist Lexicon**, zakladnica besedišča, kjer so našteti posamezni izrazi skupaj z oblikoslovnimi značilnostmi, besednimi oblikami in lemami,
- **Metathesaurus**, metatezaver, osrednji terminološki vir, ki združuje medicinske izraze iz prek 100 različnih virov in 17

[a] http://www.nlm.nih.gov/research/umls

[b] http://www.muchmore.dfki.de

jezikov. Vsakemu terminu je prirejena pojmovna koda *(CUI – Concept Unique Identifier)*, ki različne jezikovne in terminološke variante povezuje v skupni pojem,

- **Semantic Network**, pomenska mreža, ki pojme združuje v 134 pomenskih kategorij *(TUI – Type Unique Identifier)* in med njimi opredeljuje 54 možnih pomenskih razmerij.

```
<umlsterms>
        <umlsterm id="t1" from="w5" to="w5">
                <concept id="t1.1" cui="C0019134" preferred="Heparin" tui="T118 T121 T123">
                        <msh code="D9.203.698.373.400" />
                </concept>
        </umlsterm>
        <umlsterm id="t2" from="w11" to="w11">
                <concept id="t2.1" cui="C0033107" preferred="prevention & control" tui="T170" />
        </umlsterm>
        <umlsterm id="t3" from="w13" to="w13">
                <concept id="t3.2" cui="C0039798" preferred="therapeutic aspects" tui="T169" />
        </umlsterm>
        <umlsterm id="t4" from="w16" to="w16">
                <concept id="t4.1" cui="C0009566" preferred="Complication" tui="T046" />
        </umlsterm>
</umlsterms>
<semrels>
        <semrel id="r1" term1="t1.1" term2="t4.1" reltype="diagnoses" />
        <semrel id="r2" term1="t4.1" term2="t1.1" reltype="produces" />
        <semrel id="r3" term1="t1.1" term2="t4.1" reltype="affects" />
</semrels>
```

**Slika 1** Pomensko označeno besedilo.

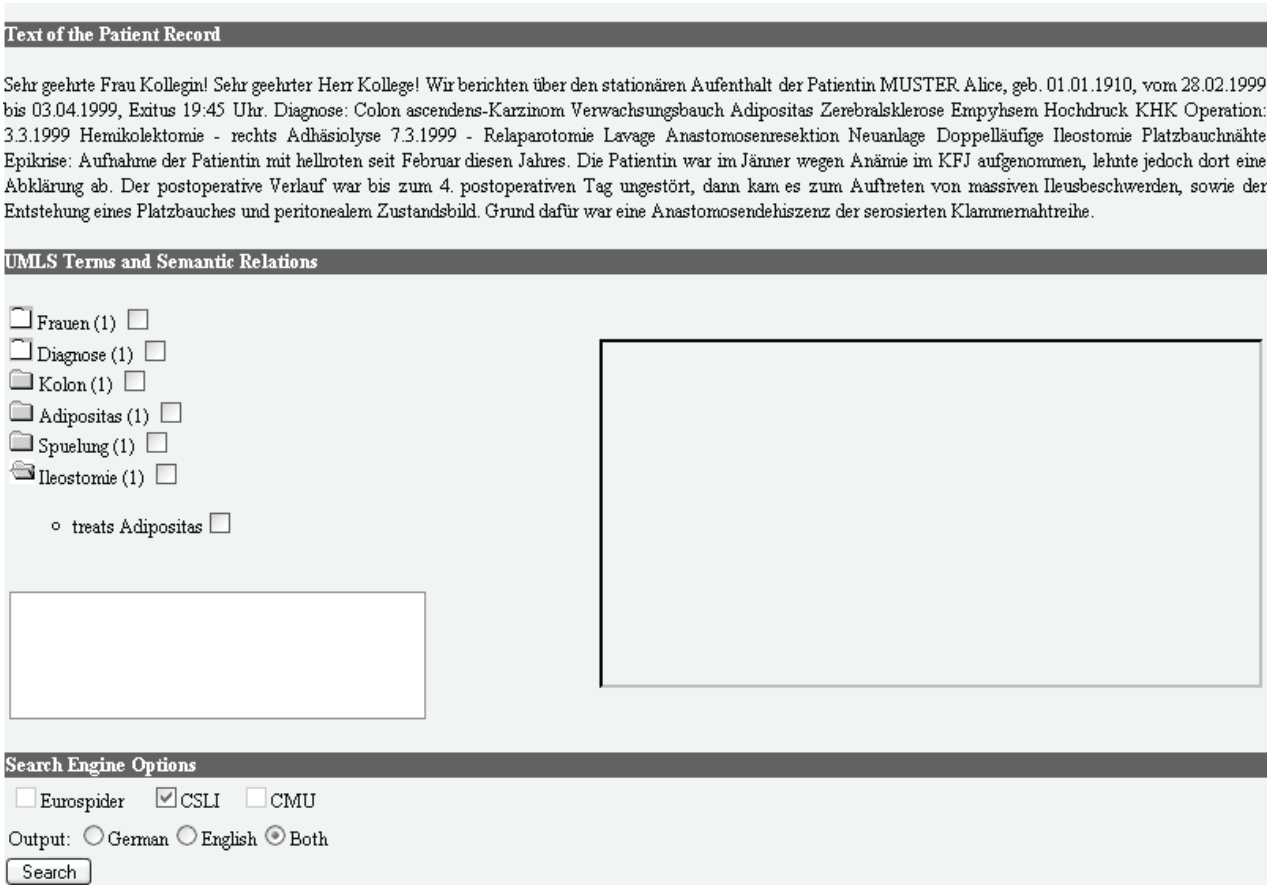
Cilj projekta MUCHMORE je bil preveriti učinkovitost medjezičnega iskanja s pomočjo pomenskega označevanja in rezultate primerjati z drugimi uveljavljenimi metodami. Preskus smo izvajali za omejeno zbirko dokumentov, in sicer 9.000 povzetkov medicinskih člankov v angleškem in nemškem jeziku, ki smo jih pridobili s Springerjevega spletišča.[c]

Za jezikovno predobdelavo besedil smo uporabili več med seboj povezanih orodij, in sicer SPPC[1] za tokenizacijo oziroma razčlembo besedila na stavke, besede in ločila, TnT[2] za besednovrstno označevanje, Mmorph[3] za oblikoslovno analizo in Chunkie[4] za razpoznavanje besednih zvez.

Sledi pomensko označevanje, kjer se na ravni termina v besedilo vnesejo naslednji podatki:

- pojmovna koda (CUI),
- koda pomenske kategorije (TUI),
- koda pripadajoče kategorije MeSH *(Medical Subject Headings)*,
- prednostni termin, tj. izraz, ki predstavlja priporočeno poimenovanje določenega pojma *(Preferred Term)*.

[c] http://www.springerlink.de




**Slika 2** Prikaz iskalne zahteve z označenimi pojmi in razmerji.

Iskanje terminov v besedilu in označevanje s podatki iz UMLS-a temelji na oblikoslovni analizi, tako da se iščejo leme in ne besedne oblike. Prav tako se poiščejo večbesedni termini in termini, ki so skriti v zloženkah, kar je še posebej pomembno za nemščino. Sistem razpoznava tudi nekatere terminološke variacije, kot so zamenjan vrstni red besed v terminu ali tipične okrajšave. Poleg tega se v besedilu samodejno označijo možna pomenska razmerja med pojmi na ravni povedi, in sicer na podlagi razmerij, ki jih definira Semantic Network.

Za primer vzemimo naslednji angleški stavek: "For many decades, heparines have been used successfully for prophylaxis and treatment of thromboembolic complications world-wide".

V tej povedi najdemo medicinske izraze *heparines, prophylaxis, treatment, thromboembolic, complications*, ki jih označimo s ustreznimi podatki iz UMLS-a. Nato preverimo vse možne kombinacije med najdenimi pomenskimi kategorijami (v tem primeru *T118, T121, T123, T170, T169* in *T046*) in ugotovimo, da med pojmoma *C0019134 Heparin* in *C0009566 Complication* obstajata možni razmerji *diagnoses* in *affects*, ki ju prav tako označimo. Zapis teh podatkov v XML obliki kaže Slika 1.

Pri samodejnem pomenskem označevanju prihaja do dvoumnosti na več ravneh. Tako lahko isti termin priredimo več različnim pojmom, določeni pojmi pa imajo lahko več možnih pomenskih kategorij, kot je razvidno iz primera *heparin*. Prav tako je lahko dvoumna preslikava izbranega pojma v hierarhično strukturo MeSH, denimo *anorexia*, ki se ji lahko priredi oznaka *C23.888.821.108* pod nadpomensko *Signs and Simptoms, Digestive*, ali *F03.375.050* pod nadpomenkama *Mental Disorders – Eating Disporders*.

Za preslikavo pojmov v strukturo MeSH smo se odločili predvsem z vidika pomenskih razmerij, saj so pomenske kategorije (TUI) in razmerja, ki jih opredeljuje UMLS, zelo splošni in zato pogosto preširoki za namene medjezičnega iskanja. Poleg tega smo nameravali v besedilih odkrivati nova pomenska razmerja, za to pa je drevesna struktura MeSH primernejša, saj je pri vsakem pojmu možno izbrati bolj ali manj specifično pomensko kategorijo.

## Prototip medjezičnega iskalnika s pomočjo pojmov

Naša ciljna aplikacija je bila iskalnik po zbirki angleških in nemških medicinskih člankov, kjer je tipična iskalna zahteva v obliki elektronske pacientove kartoteke oziroma zdravnikove zabeležke o pregledu. Sistem je namenjen predvsem nemško govorečim uporabnikom, ki torej iskalno zahtevo sprožijo v nemščini, a jih zanima strokovna literatura v obeh jezikih. Preskusna različica je na voljo na naslovu http://muchmore.dfki.de.

Iskalna zahteva se sproti obdela z jezikovnega vidika, tako da je omogočeno boljše prepoznavanje terminov, nato pa sledi prej opisano pomensko označevanje. Po pomenski analizi sistem uporabniku ponudi pojme in možna razmerja med njimi, s katerimi lahko svojo medjezično iskalno zahtevo natančneje oblikuje. Če sistem v iskalni zahtevi kakega izraza ni našel, ga lahko uporabnik doda ročno v spodnje okence. Nato se sproži iskanje po zbirki dokumentov, ki prek pomenskih oznak poišče angleške in nemške zadetke.

### Rezultati

Da bi ugotovili, ali pojmovni pristop v medjezičnem iskanju deluje bolje kot druge metode, smo izvedli niz preskusov. Za to smo uporabili že omenjeno zbirko medicinskih povzetkov v angleškem in nemškem jeziku in seznam 25 iskalnih zahtev, ki so nam jih posredovali medicinski strokovnjaki. Da lahko ocenjujemo kakovost sistema, moramo poznati pravilni izbor dokumentov za vsako iskalno zahtevo. Ta izbor so prav tako opravili medicinski strokovnjaki, skupno število izbranih dokumentov pa je bilo 959. To število v nadaljevanju predstavlja 100-odstotni priklic, če poženemo vseh 25 iskalnih zahtev. Testne iskalne zahteve so večinoma kratke in jedrnate, denimo "Atrthroscopic treatment of cruciate ligament injuries" ali "Indication for implantable cardioverter defibrillator (ICD)".

Čeprav smo imeli na razpolago vzporedni angleško-nemški korpus, se pri evaluaciji medjezičnega iskanja pretvarjamo, da temu ni tako. Namesto tega smo poskušali prek iskalnih zahtev v nemščini dostopati do angleških dokumentov.

Najprimitivnejši pristop k medjezičnemu iskanju je, da z besedami nemške iskalne zahteve skušamo neposredno poiskati angleške dokumente. Utemeljitev tega pristopa je, da bo že spričo delnega prekrivanja strokovne terminologije v obeh jezikih mogoče najti ustrezne dokumente. Zares se pokaže, da na ta način "uganemo" 66 ustreznih dokumentov (glej vrstico DE2EN-token v Tabeli 1). Najbolje so se odrezale iskalne zahteve, ki so vsebovale kratico HIV ali latinske izraze, npr. diabetes mellitus.

Kot drugo primerjavo smo preskusili strojno prevajanje iskalnih zahtev iz nemščine v angleščino s komercialnim prevajalnikom PersonalTranslator 2001 (Linguatec, München). Čeprav je bilo v prevajalniku mogoče izbrati strokovno besedišče medicine in kemije, se je pokazalo, da program mnogih medicinskih izrazov ni zmogel prevesti, ker jih ni imel v slovarju. Pri drugih izrazih prevod ni bil mogoč, ker program ne zna analizirati zloženk. Tako se pri besedi Myokardinfarkt tudi Infarkt ni prevedel, čeprav ga ima v slovarju. Spet druge iskalne zahteve so se s strojnim prevajalnikom prevedle brezhibno, denimo:

- DE: Möglichkeiten der Korrektur von Deformitäten in der Orthopädie

- EN: Possibilities of the correction of deformities in orthopedics.

Vrstica DE2EN-MT v tabeli kaže, da smo s tem pristopom našli 376 ustreznih dokumentov.

Nato smo te rezultate primerjali z rezultati, ki jih dobimo, če medjezikovno povezavo ustvarimo s pomočjo pojmovnih oznak iz UMLS-a. V teh poskusih se torej v nemški iskalni zahtevi poiščejo vsi medicinski termini, ki se jim priredijo pojmovne oznake iz Metathesaurusa (DE2EN-CUI), pojmovne oznake iz hierarhije MeSH (DE2EN-MeSH), na koncu pa se poiščejo še pomenska razmerja med njimi (DE2EN-SemRel). Prek teh oznak se išče ustrezne dokumente v angleški zbirki. Vrstica DE2EN-all-sem v tabeli kaže rezultat, ki ga dobimo, če pri iskalni zahtevi upoštevamo vse pomenske informacije skupaj. Kot je razvidno, pri slednjem dobimo tudi najboljši rezultat v sklopu pojmovno usmerjenih preskusov.[5]

**Tabela 1** Rezultati medjezičnega iskanja.

| | mAvP | št.dok. | AvP 0.1 | P10 |
|---|---|---|---|---|
| DE2EN-token | 0.0512 | 66 | 0.1530 | 0.1160 |
| DE2EN-MT | 0.1184 | 376 | 0.3382 | 0.2520 |
| DE2EN-CUI | 0.1620 | 366 | 0.3724 | 0.2800 |
| DE2EN-MeSH | 0.1699 | 304 | 0.3888 | 0.2600 |
| DE2EN-SemRel | 0.0229 | 23 | 0.0657 | 0.0480 |
| DE2EN-all-sem | 0.1774 | 404 | 0.3872 | 0.2720 |
| DE2EN-SimThes | 0.2290 | 409 | 0.4492 | 0.3640 |
| DE2EN-all | 0.2955 | 518 | 0.5761 | 0.4600 |

Da bi bila slika popolna, smo preskusili še tretjo metodo, ki na področju iskanja podatkov velja za najbolj obetavno, in sicer slovar podobnih izrazov (Similarity Thesaurus). Ta vsebuje samostalnike, samostalniške zveze, pridevnike in glagole, ki se v našem korpusu najpogosteje pojavljajo v sorodnih kontekstih in tako sklepamo, da imajo soroden pomen. Tak slovar je mogoče zgraditi tudi iz enojezičnega korpusa in ga uporabljati pri enojezičnem iskanju podatkov, v našem primeru pa smo ga zgradili na podlagi vzporednega korpusa. Tako pridobljeni slovar podobnih izrazov je za vsako nemško besedo vseboval množico najverjetnejših prevodnih ustreznic in ostalih sorodnih besed. Te angleške besede so se uporabile za iskanje ustreznih angleških dokumentov. Spodnji primer kaže seznam angleških besed, ki jih slovar podobnih izrazov navaja za nemški Myokardinfarkt: infarction, acute myocardial infarction, mycardial, thrombolytic, acute, thrombolysis, crs, synchronisation, cardiogenic shock, ptca.

Poleg števila ustreznih dokumentov, ki ga navaja tretji stolpec Tabele 1, merimo učinkovitost našega sistema z ustaljenimi metodami projektov TREC.[6] Drugi stolpec tako vsebuje splošno natančnost sistema, ki se računa kot srednja povprečna natančnost (*mean average precision* - mAvP). Natančnost pri tem pomeni razmerje med pravilno izbranimi dokumenti in vsemi izbranimi dokumenti. Predzadnji stolpec kaže povprečno natančnost ob priklicu 0,1 (AvP 0,1), saj Eichmann in soavtorji[7] ugotavljajo, da je učinkovitost sistema bolje meriti v območju visoke natančnosti, s predpostavko, da uporabnike predvsem zanimajo visoko uvrščeni zadetki. Zadnji stolpec v skladu s tem navaja natančnost za 10 najvišje uvrščenih dokumentov (P10).

Kot komentar k prikazanim rezultatom povejmo, da nas je v okviru opisanih eksperimentov zanimalo predvsem, ali je s pomočjo pomenskega označevanja mogoče prekositi doslej uveljavljene metode s strojnim prevajanjem in s slovarjem sorodnih besed. Pokazalo se je, da nobena od posameznih pomenskih kategorij (CUI, MeSH ali SemRel) ni dovolj zanesljiva, da bi sama po sebi pomagala izbrati ustrezne dokumente v drugem jeziku, pač pa se zelo dobro odreže kombinacija vseh pomenskih kategorij (all-sem).

Metoda s slovarjem sorodnih besed sicer da boljši rezultat (SimThes), vendar jo moramo upoštevati z rezervo, saj smo slovar zgradili iz istega vzporednega korpusa, na katerem so potekali preskusi. Tako je bil slovar pravzaprav preveč ukrojen po meri, tako da v realnih situacijah medjezičnega iskanja takih rezultatov ne bi mogli ponoviti. Zanimivo je, da kombinacija slovarja sorodnih besed s pomenskimi kategorijami da najboljše rezultate od vseh (all), kar znova potrjuje

koristnost slednjih. Seveda pa se je ob tem treba zavedati, da je področje medicine izjemno v smislu opremljenosti s terminološko-ontološkimi viri – čeravno tu slovenščina še precej zaostaja – in da si podobnih aplikacij na drugih področjih ne moremo zamišljati brez predhodnega vložka v (nad)gradnjo teh virov.

## Odkrivanje pomenskih razmerij v medicini

V nadaljevanju prispevka se posvečamo drugi, bolj raziskovalno usmerjeni aplikaciji pomenskega označevanja, in sicer odkrivanju novega znanja in novih povezav med znanimi dejstvi. Tu smo se posebej osredotočili na pomenska razmerja na področju medicine, pri čemer se je motivacija za pričujočo raziskavo porodila iz opažanj, da projekcija pomenskih razmerij iz UMLS-a na besedilo, se pravi tistih, ki jih med pomenskimi kategorijami (*Semantic Type*) opredeljuje Semantic Network, redko ustreza dejanskim pomenskim razmerjem, ki so udejanjena v besedilu.

Razlogov za to je več. Semantic Network definira skupno 54 pomenskih razmerij, specifičnih za področje medicine, kot so *location_of, affects, treats, causes, interacts_with*. Ta razmerja se ne nanašajo na konkretne pojme, ampak so opredeljena na ravni pomenskih kategorij, na primer *Pharmacologic Substance - affects - Cell Function*. Če ta zelo splošna razmerja, ki seveda držijo le za omejeno podmnožico družine farmakoloških pripravkov in celičnih funkcij, avtomatsko preslikamo na dokumente, številna ugotovljena razmerja ne držijo. Tako se denimo v stavku pojavita pojma *discectomy* in *history*, ki ju uvrstimo v pomenski kategoriji *Therapeutic Procedure* in *Occupation or Discipline*. Med tema dvema kategorijama UMLS definira razmerje *method_of*, vendar v našem primeru enačba *discectomy* = *method_of history* ne drži.

Da je stvar še bolj meglena, sta dve pomenski kategoriji teoretično lahko povezani s številnimi različnimi, celo nasprotnimi razmerji, kot kaže spodnji primer:

- Therapeutic Procedure | **prevents** | Neoplastic Process
- Therapeutic Procedure | **complicates** | Neoplastic Process
- Therapeutic Procedure | **affects** | Neoplastic Process
- Therapeutic Procedure | **treats** | Neoplastic Process
- Therapeutic Procedure | **associated_with** | Neoplastic Process

Želeli smo torej ugotoviti, ali je s pomočjo vseh ostalih podatkov, ki jih imamo na razpolago, se pravi jezikoslovne analize, pojmovnih oznak (CUI) in oznak iz MeSH-a, mogoče razlikovati dejanska razmerja od hipotetičnih in ali se dejanska razmerja v besedilih izražajo na ustaljene načine.

### Sorodne raziskave

Da je s pomočjo statistične obdelave medicinskih besedil mogoče priti do novih in koristnih hipotez, je prvi pokazal že Swanson.[8] Njegove zamisli so naprej razvijali Weeber in soavtorji,[9] ki opisujejo naprednejši model obdelave medicinskih besedil z jezikovno in pomensko komponento. Večina raziskovalcev se s to temo ukvarja v zvezi s samodejno izdelavo ontologij, zato se v besedilih predvsem išče taksonomske povezave med pojmi, se pravi razmerja hipo- in hipernimije. Maedche in Staab,[10] denimo, razvijata metodo učenja asociativnih pravil za iskanje pojmovnih mrež. Za razliko od omenjenih pristopov se je naša raziskava usmerila v odkrivanje znanih, za medicino tipičnih razmerij, vendar brez uporabe razmerij, ki jih opredeljuje Semantic Network.

### Razlikovanje med dejanskimi in hipotetičnimi razmerji

Cilj postopka je presejati samodejno označena pomenska razmerja, ki bi lahko držala med pomenskimi kategorijami, in izmed njih izbrati tista, ki so v besedilu res izražena. Izhajamo iz dveh predpostavk:

- pomenska razmerja se v besedilu udejanjajo z določenimi jezikovnimi sredstvi, predvsem z glagoli, in
- pomenska razmerja so bolj pomembna, če se pojavijo med dvema pomembnima, tj. čimbolj specializiranima, pojmoma.

V ta namen smo preverili sopojavljanje posameznih razmerij s posameznimi glagoli v korpusu in za vsak glagol zbrali seznam petih najbolj tipičnih razmerij, ki jih lahko označuje, kot kaže spodnji primer:

- activate (84)
- interacts_with (1937)
- produces (836)
- affects (544)
- disrupts (324)
- result_of (295)

Obenem smo v duhu druge predpostavke pojme obtežili s statistično vrednostjo IDF (*Inverse Document Frequency*), ki višje uvršča pojme, ki se pojavljajo le v omejenem številu dokumentov, nižje pa splošne pojme, kot so *patient, therapy, result* itd. Obe metodi smo združili in uporabili kot sito za izbiro dejanskih in pomembnih razmerij. Po uporabi obeh sit v besedilu ostane približno 31% samodejno označenih razmerij.

### Odkrivanje novih pomenskih razmerij

Ob pregledovanju samodejno označenih pomenskih razmerij pred in po situ smo opazili, da številna razmerja še vedno ostajajo neopažena, ker niso zapisana v UMLS-u. Poleg tega se za namene odkrivanja znanja pokaže, da je 54 pomenskih razmerij, kot jih opredeljuje Semantic Network, pravzaprav preveč, saj so si številna med njimi zelo podobna (npr. *interacts_with* in *associated_with*). Tako smo se odločili, da namesto vseh 54 razmerij poskušamo odkrivati le 15 najpogostejših.

Izhajamo iz predpostavke, da pogosto sopojavljanje dveh medicinskih pojmov v besedilu pomeni, da sta pojma med seboj nekako povezana. Če bi skušali ugotavljati pomensko razmerje za vsak par pojmov posebej, bi imeli na eni strani preveč takih pojmovnih parov, na drugi strani pa premajhne pogostosti njihovih sopojavitev, da bi na tej podlagi lahko ugotovili pomensko razmerje. Pojme zato prevedemo na njihovo splošnejšo obliko, in sicer kategorijo MeSH drugega reda.

V drevesni strukturi MeSH-a so osnovne veje označene s črkami, npr. *A - Anatomy, B - Organisms, C - Diseases*. Na naslednji stopnji se ta delijo v bolj specializirane skupine pojmov, denimo *B05 - Fungi ali C02 - Virus Diseases*. Predpostavljamo torej, da pogosto sopojavljanje dveh podkategorij MeSH (npr. *A01|C23*) pomeni, da med njima obstaja pomensko razmerje.

Za vsako od izbranih 15 razmerij smo izdelali seznam 300 najpogostejših parov podkategorij MeSH:

- treats – D27|C23, D3|C23, E7|C23, E7|C2 …

Nato v korpusu označimo vsa domnevna razmerja, ki ustrezajo temu seznamu, izločimo pa tista, ki vključujejo preveč splošne termine.

Uspešnost te metode smo nato preskusili na dva načina. Pri prvem nas je zanimalo, kako izboljšanje kakovosti pomenskih razmerij oziroma dodajanje

novih razmerij vpliva na pojmovno iskanje podatkov.

Izdelali smo pet različic korpusa in iskalnih zahtev. Pri prvi smo iz obstoječih samodejno označenih pomenskih razmerij odstranili vsa tista, ki so vključevala preveč splošne pojme (UMLS-IDF), pri drugi smo poleg prvih odstranili tudi vsa razmerja, ki niso vsebovala tipičnih glagolov (UMLS-IDF-glag). Nato smo v tako prečiščeni korpus dodali novo odkrita razmerja (UMLS-IDF-glag), poleg tega pa smo novo odkrita razmerja dodali tudi prvotnemu neprečiščenemu korpusu (UMLS+nove). Nazadnje smo korpus označili le z novo odkritimi razmerji (samo nove).

**Tabela 2** Evalvacija izbiranja in odkrivanja pomenskih razmerij z medjezičnim iskanjem.

| | mAvP | št. dok. | AvP 0.1 | P10 |
|---|---|---|---|---|
| UMLS-IDF | 0,126 | 203 | 0,315 | 0,280 |
| UMLS-IDF-glag | 0,107 | 175 | 0,282 | 0,264 |
| UMLS-IDF-glag+nove | 0,124 | 197 | 0,336 | 0,308 |
| **UMLS+nove** | **0,153** | **259** | **0,419** | **0,344** |
| samo nove | 0,116 | 213 | 0,363 | 0,280 |

Rezultati v Tabeli 2 so za medjezično iskanje seveda vsi po vrsti slabi, vendar jih moramo razumeti zgolj kot preskus uspešnosti naše metode izbiranja in odkrivanja pomenskih razmerij. Iz njih precej jasno izhaja, da najboljši rezultat dobimo, če uporabimo čimveč podatkov, se pravi razmerja iz UMLS-a in zraven še novo odkrita razmerja. To je v skladu z ugotovitvami številnih avtorjev,[11] da je pri iskanju podatkov širjenje iskalne zahteve skoraj zmeraj uspešnejša strategija od oženja ali natančnejše opredelitve iskalne zahteve, čeprav se to morda na prvi pogled ne zdi logično.

Druga faza evalvacije je zajemala primerjavo naših metod sejanja in odkrivanja razmerij z besedili, kjer so pojme in razmerja ročno označili medicinski strokovnjaki. Žal smo imeli na razpolago omejeno število takšnih besedil, nadaljnja težava pa je bila dejstvo, da se pri svojem označevanju strokovnjaki niso omejili na istih 15 razmerij. Kljub temu primerjava našega prečiščenega in obogatenega korpusa UMLS-IDF-glag+nove kaže 63% ujemanje z ročno označenimi razmerji, kar je spodbuden rezultat.[12]

## Razprava

Samodejna pomenska analiza besedil s pomočjo ontologije, ki jo opisujemo v pričujočem prispevku, je pomembna raziskovalna veja ne le v kontekstu medjezičnega iskanja, ampak tudi kot pomembna tehnologija v okviru semantičnega spleta. Pri opisani raziskavi in predstavljenih poskusih pa smo imeli v prvi vrsti pred očmi uporabniški vidik, se pravi omogočanje učinkovitejšega dostopa do relevantnih informacij prek jezikovnih meja.

Medtem ko je mogoče trditi, da so rezultati našega pojmovno obogatenega medjezičnega iskanja na medicinskem področju vidna izboljšava dosedanjih zabeleženih rezultatov, so izsledki pri odkrivanju pomenskih razmerij šele na poti k dejanski uporabnosti. Nadaljnje raziskave vodijo predvsem k metodam strojnega učenja in vizualizacije podatkov.[13]

Da je pojmovno obogateno medjezično iskanje dokumentov za uporabnike izredno dobrodošla in koristna aplikacija, se je pokazalo v fazi zaključne evalvacije projekta Muchmore, kjer so iskalnik ocenjevali angleško in nemško govoreči medicinski strokovnjaki.[14] Povprečna skupna ocena iskalnika je dosegla 3,40 (na lestvici od 1 do 5, kjer je 5 najvišja ocena), vendar so se uporabniki strinjali, da je način iskanja na podlagi pacientove kartoteke in pojmovne analize izredno zanimiv in bi v prihodnosti utegnil nadomestiti klasično iskanje s ključnimi besedami. Po drugi strani pa nedvomno drži, da bo do prave premostitve jezikovnih preprek moralo preteči še precej raziskovalne vode. Še posebej, če naj bi rešitve vključevale tudi slovenščino.

**Literatura**


1. Piskorski J, Neumann G: An intelligent text extraction and navigation system. Proceedings of the 6th RIAO, Paris, 2000.
2. Brants T: TnT – A Statistical Part-of-Speech Tagger. Proceedings of the 6th ANLP Conference, Seattle, WA, 2000.
3. Petitpierre D, Russel G: MMORPH - The Multext Morphology Program. Multext deliverable report for the task 2.3.1, ISSCO, University of Geneva, Switzerland, 1995.
4. Skut W, Brants T: A maximum entropy partial parser for unrestricted text. Proceedings of the 6th ACL Workshop on Very Large Corpora (WVLC), Montral, Canada, 1998.
5. Volk M, Ripplinger B, Vintar Š, Buitelaar P, Raileanu D, Sacaleanu B: Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval. International Journal of Medical Informatics, Volume 67:1-3, December 2002.
6. Gaussier E, Grefenstette G, Hull DA, Schulze BM: Xerox TREC-6 site report: Cross language text retrieval. Proceedings of the Sixth TExt Retrieval Conference (TREC-6). National Institute of Standards Technology (NIST), Gaithersburg, MD, 1998.
7. Eichmann D, Ruiz M, Srinivasan P. Cross-Language Information Retrieval with the UMLS Metathesaurus. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998.
8. Swanson DR, Smalheiser NR: An interactive system for finding complementary literatures: A stimulus to scientific discovery. Artificial Intelligence 1997, 91:183-203.
9. Weeber M, Klein H, Aronson JG, Mork L, Jong van den Berg, Vos R: Text-Based Discovery in Biomedicine: The architecture of the DAD-System. The American Medical Informatics Association 2000 Symposium.
10. Maedche A, Staab S: Discovering Conceptual Relations from Text. W. Horn (ed.) ECAI 2000, Proceedings of the 14th European Conference on Atrificial Intelligence, Berlin, IOS Press, Amsterdam.
11. Schäuble P, Sheridan P: Cross-language information retrieval (CLIR) track overview. Proceedings of the Sixth TExt Retrieval Conference (TREC-6). National Institute of Standards Technology (NIST), Gaithersburg, MD, 1998.
12. Vintar Š, Buitelaar P, Volk M: Semantic Relations in Concept-Based Cross-Language Medical Information Retrieval. Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining (ATEM), 2003, Cavtat-Dubrovnik, Croatia.
13. Vintar Š, Todorovski L, Sonntag D, Buitelaar P: Evaluating Context Features for Medical Relation Mining. Proceedings of the European Workshop on Data Mining and Text Mining for Bioinformatics, held in conjunction with ECML/PKDD, 2003, Dubrovnik, Croatia.
14. Report on User Evaluation, Deliverable 10.1 of the IST-1999-11438 project Muchmore, http://muchmore.dfki.de.