Razreševanje sklicev pri analizi slovenskih besedil

Peter Holozan

Amebis, d. o. o.
Bakovnik 3, 1241 Kamnik
peter.holozan@amebis.si

Povzetek

Razreševanje sklicev je pomemben del jezikovnih tehnologij, vendar za slovenščino ta tehnologija še ni bila razvita. Obstajajo različne vrste sklicev, članek se osredotoča predvsem na anafore pri osebnih zaimkih. Uporabljene so bile štiri metode razreševanja, ki se med seboj dopolnjujejo, najpomembnejša temelji na metodah na osnovi aktivacije. Prvi rezultati so obetavni, razreševanje sklicev je bilo uporabljeno tudi v sistemu za odgovarjanje na vprašanja Piflar, ki zna s tem odgovoriti na več vprašanj.

Reference Resolution for Slovenian Texts Analysis

Reference resolution is an important part of language technologies, but has not yet been developed for Slovenian. There are various types of references and the paper focuses on anaphora resolution of personal pronouns. Four methods, used in combination, were used; the most important one is based on activation. First results are promising: reference resolution was used in the question answering system Crammer, which can, as a result, answer more questions than before.

1. Uvod

Razreševanje sklicev je pomemben del jezikovnih tehnologij. Sklice lahko razdelimo na anafore, kjer je razlaga pred sklicem, in katafore, kjer je razlaga za sklicem. Tipičen primer sklicev so zaimki, niso pa sklici omejeni le na zaimke, čeprav se največkrat omejimo le nanje.

Razreševanje sklicev lahko pomaga tudi pri razdvoumljanju besedil, saj v precej primerih tega ne moremo narediti brez razreševanja sklicev, iz česar sledi, da v resnici ne moremo izhajati iz tega, da imam vhodno besedilo že razdvoumljeno, ampak se morata razdvoumljanje in razreševanje sklicev dopolnjevati. (McShane, Beale, Nirenburg, 2010) Tak primer sta npr. povedi »Miha je videl matico, ki jo je privil Janez.« in »Miha je videl matico, ki jo je vzgojil Janez..«, kjer je pomen besede »matica« odvisen od prilastkovega odvisnika, kjer je »matica« nadomeščena z osebnim zaimkom »jo«.

V veliko primerih se razreševanje omeji le na razreševanje anafor. (Mitkov, 1999; Němčík, 2006)

Za slovenščino razreševanje sklicev še ni bilo narejeno, zato je smiselno preizkusiti, kako uspešno se to da vgraditi v analizator, ki prevaja naravni jezik v Amebisov vmesni jezik, katerega podrobni opis je v prilogi 6.2 v (Holozan, 2011). Ta vmesni jezik uporabljajo mnogi izdelki podjetja Amebis, npr. strojni prevajalnik Presis in sistem za odgovarjanje Piflar, kar pomeni, da bo ta izboljšava vplivala tudi nanje.

Najprej bo v razdelku 2 predstavljen problem sklicev, pri čemer bodo omenjene razlike pri sklicih med slovenščino in angleščino, za katero je bilo opravljeno največ dela pri razreševanju sklicev.

V razdelku 3 bodo opisane nekatere obstoječe metode razreševanja sklicev.

Nato bodo v razdelku 4 preizkušene različne metode za razreševanje sklicev, ki se izvajajo zaporedno in iz katerih se dobi skupni rezultat označevanja referenc.

Na koncu bodo v razdelku 4 predstavljeni rezultati, v razdelku 5 pa bo razreševanje sklicev uporabljeno v sistemu za odgovarjanje na vprašanja v naravnem jeziku Piflar.

2. Sklici

Glavni kandidati za sklice so zaimki, in sicer predvsem osebni. V Gigafidi pokrivajo osebni zaimki 1,4 % vseh besednih pojavnic, vendar je treba upoštevati še, da v slovenščini velik del osebnih zaimkov izpuščamo (v povedih iz korpusa jos100k (Erjavec, Krek, 2008), ki jih je analizatorju uspelo analizirati, je izpuščenih osebnih zaimkov več kot dvakrat toliko kot neizpuščenih), jih je pa vseeno treba razrešiti, če želimo dobiti pravi pomen teh stavkov.

V avtodomu sta ***onadva*** prevažala opij

Policisti in kriminalisti so v sodelovanju s cariniki na avtocestnem počivališču v bližini Murske Sobote zaustavili tovorni avtomobil, ki je bil predelan v avtodom. Po pregledu avtomobila so *oni* v njem našli 390 gramov opija, tri vrečke obrezanih makovih glavic in tri grame halucinogenih gobic. Vozilo je bilo registrirano v Franciji. V njem pa sta se vozila **Francoza**, stara 32 in 33 let, so sporočili *oni* s policijske uprave v Murski Soboti.

Zaradi utemeljenega suma neupravičene proizvodnje in prometa s prepovedanimi drogami sta bila **tujca** s kazensko ovadbo privedena pred preiskovalnega sodnika okrožnega sodišča v Murski Soboti, ki je zoper **oba** odredil pripor.

Slika 1. Primer.

Slika 1 prikazuje primer besedila s sklici. V kurzivi so dodani sicer izpuščeni osebni zaimki, odebeljeno pa so označeni sklici, ki jih je treba povezati. Podobno lahko povežemo še »avtodomu«, »tovorni avtomobil«, »ki«, »avtomobila«, »vozilo« in »njem«.

Razreševanje sklicev, ki niso zaimki, je v slovenščini težje kot v angleščini, ker slovenščina ne uporablja členov, zato se iz tega, da ima neka samostalniška fraza določni člen, ne da sklepati na to, da se nanaša na nekaj, kar je bilo omenjeno že prej, kot je to primer v angleščini. To pomeni, da se moramo v slovenščini tukaj veliko bolj

zanesti na pomene, delno pa tudi na besedni red (členjenje po aktualnosti).

Težava so tudi sklici, ki so v angleščini zapisani z besedo »one«, npr.: »If you cannot attend a tutorial in the morning, you can go for an afternoon one.« (Mitkov, 1999). Slovenski prevod bi bil: »Če se ne moreš udeležiti vaj dopoldne, greš lahko na popoldanske.« V slovenščini tukaj ni posebne besede, na katero bi lahko vezali sklic, ampak bi tukaj lahko rekli, da gre za izpust besede »vaje« v drugem stavku.

Sklici lahko povezujejo več predhodnih besed v eno besedo ali obratno. V primeru »Srečal sem Johna in Mary. Bila sta zelo vesela, saj smo dobri prijatelji.« se John in Mary najprej povežeta v izpuščeni zaimek »onadva«, nato pa še skupaj s pripovedovalcem (1. osebo) v izpuščeni zaimek »mi«. Obratno pa je v primeru »Starejši par je hodil po parku in moški se je nenadoma spotaknil.«, ko je »moški« najverjetneje del »para« iz predhodnega stavka.

Razreševanje sklicev je zelo odvisno od pomenov. Če uporabim primer iz (Němčík, 2006): »John je skril Billove ključe. Bil je pijan.«, se ljudem zdi najverjetnejša interpretacija, da se drugi stavek nanaša na Billa, ker pač sklepamo, da je vožnja pod vplivom alkohola nevarna in je Johna skrbelo za Billa, zato mu je skril ključe, da ne bi mogel odpeljati. Ni pa to edina možna interpretacija, morda je John bil pijan in je hotel nagajati Billu in mu je zato skril ključe hiše. Taki primeri kažejo na to, da je razreševanje sklicev res zahteven problem za računalnike.

3. Nekatere metode razreševanja sklicev

Za razreševanje sklicev je bila razvita množica metod in nekatere bodo na kratko predstavljene v nadaljevanju tega razdelka.

3.1. Hobbsovo sintaktično iskanje

Hobbsovo sintaktično iskanje (Hobbs, 1978) je bila prva metoda, ki je uporabila jezikovno znanje in je kljub starosti in relativni preprostosti (že sam Hobbs je menil, da je to le naivna metoda) še vedno primerljivo uspešna v primerjavi z modernejšimi metodami (Němčík, 2006).

Osnova za postopek je drevo izpeljav za poved. Hobbsovo iskanje določi vrstni red, v katerem samostalniške fraze postanejo kandidate za razreševanje sklicev. V drevesu začnemo iskati levo od zaimka, za katerega želimo razrešiti sklic, potem pa se dvigamo in vsakič iščemo v širino od leve proti desni.

Metodo lahko dopolnimo s pomenskimi omejitvami pri kandidatih.

Težava pri metodi je, da lahko vedno najdemo primere, v katerih ne deluje, dodatno pa je izdelava drevesa izpeljav sama po sebi zapleten problem.

3.2. Algoritem BFP

Algoritem BFP (Brennan, Friedman, Pollard, 1987) temelji na teorija fokusa (centering theory), ki je bila prvič opisana v (Joshi, Kuhn, 1979). Ta opisuje, kako se spreminja fokus diskurza, ena od metod fokusiranja pa je tudi uporaba zaimkov, ki nas usmerjajo na fokus. Ta se lahko spreminja z različnimi vrstami prehodov.

Pokazalo pa je se, da razvoj v smeri vedno bolj kompleksnih pravil slepa ulica, ker ni bilo mogoče dovolj podrobno zajeti splošnega znanja in opisati jezika, zato so se metode usmerile v smeri, ki zahteva manj znanja (Němčík, 2006).

3.3. Faktorji poudarka

Postopek s faktorji poudarka (salience factors) je bil predlagan v (Lappin, Leass, 1994). Ti faktorji so uteži, ki so prirejene posamičnim možnostim sklicev in potem kombinirane, da se določi najpomembnejši element diskurza. Dodatno postopek ugotavlja, kateri zaimki so del fraz in nimajo sklicev (npr. »it« v »It's raining.«) in določa povratne zaimke (Němčík, 2006).

Uteži je treba določiti z eksperimentiranjem, kar pomeni, da potrebujemo korpus primerov, da lahko avtomatsko preverjamo različne uteži.

3.4. Robustni sistemi z malo potrebnega znanja

Primer za tak sistem je MARS (Mitkov, Evans, Orasan, 2002). Sistem temelji na množici predhodnostnih kazalnikov (set of antecedent indicators). Vsak od njih opisuje določen pogoj, ki se nanaša na danega kandidata za sklic, in vpliv, ki ga ima na verjetnost, da je to verjetni izvor sklica (Němčík, 2006).

Prednost te metode je, da ne potrebuje zunanjega skladenjskega razčlenjevalnika, za večino kazalnikov pa se zdi, da je jezikovno neodvisna, zato je bila te metoda uporabljena tudi za druge jezike, kot so francoščina, poljščina, arabščina in bolgarščina (Němčík, 2006).

3.5. Statistične metode

Po letu 1990 so se za razreševanje sklicev začele uporabljati tudi statistične metode (in tudi druge metode strojnega učenja). Primer je (Ge, Hale, Charniak, 1998).

Vendar vse te metode zahteva korpus učnih primerov, ki ga za slovenščino še nimamo, zato se za zdaj nismo usmerili v to smer.

4. Uporabljene metode razreševanja

Ideja razreševanja sklicev, ki jo opisujemo v nadaljevanju, je uporaba množice metod, od katerih vsaka razrešuje določene sklice, metode pa se uporabljajo od bolj proti manj zanesljivim (v tem vrstnem redu so tudi opisane, poudariti pa je treba, da je zanesljivost v tem trenutku le ocena, ki potrebuje še bolj temeljito preverjanje na večjem številu primerov).

Izbrane so bile metode, ki ne potrebujejo učnega korpusa, ker tega za slovenščino še ni. Obstaja pa po drugi strani možnost, da bi si lahko s temi za zdaj uporabljenimi metodami pomagali, da se naredi osnutek korpusa primerov sklicev, ki se potem še ročno dopolni, da ni treba celotnega izdelati ročno.

Sklici so v vmesnem jeziku opisani z novim elementom ORI, ki je dodan k obstoječemu elementu (največkrat je to osebni zaimek (OSZ) oz. navidezni (ki se skriva v osebni glagolski obliki) osebni zaimek (NOZ), lahko pa tudi drug samostalniški zaimek (SAZ) ali samostalnik (SAM)) v element JED (jedro dela samostalnike fraze) in element ORI vsebuje element SFR (samostalniško frazo). Slika 2 prikazuje primer, ko je sklic dodan osebnemu zaimku v slogi osebka (element OSB).

(1OSB:(-SFR:(-DSF:(-JED:(-OSZnemt:[10])**,(-ORI:(-SFR:(-DSF:(-JED:(-SAMe:{7d62a7;4207ac9}[/]<dc>)))))**)))))

Slika 2. Primer zapisa sklica v vmesnem jeziku

Za preizkušanje so bili uporabljeni umetno skonstruirani primeri, pravljica Rdeča kapica, Cankarjev Na klancu, testno besedilo iz priročnika Pravipis Aleksandre Kocmut, Wikipedija, šala neznanega izvora in prispevek iz črne kronike.

4.1. Izpusti osebka

To je vrsta sklicev, ki jih je mogoče zelo zanesljivo razrešiti. Gre za zaporedna stavka, pri čemer je v drugem izpuščen osebek, tako da se uporabi kar osebek iz prvega stavka: »Miha je prišel do vrat in pozvonil.«. V teh primerih se običajno izpusti še pomožni glagol, lahko pa tudi veznik: »Metka je rekla, da rada pleše in poje.«

4.2. Prilastkovi odvisniki

Tudi pri prilastkovih odvisnikih vemo, da se zaimek (»ki«, »kateri« ali pa naslonska oblika osebnega zaimka ob »ki«) v odvisniku nanaša na besedo, ki je jedro ob tem odvisniku. V primeru »Miha je videl sliko, ki jo je naslikal Janez.« tako vemo, da se »jo« nanaša na besedo »sliko«.

Težava lahko nastopi le v primerih, ko ni jasno, kaj je jedro: »Bila sta privedena pred preiskovalnega sodnika okrožnega sodišča v Kamniku, ki je zoper oba odredil pripor.« V takih primerih se lahko zgodi, da analizator označi kot jedro »Kamnik«, kar morajo potem razrešiti pomenske omejitve.

4.3. Delna osebna imena

Še posebej v časopisnem poročanju je običajno, da se oseba prvič navede s polnim imenom, v nadaljevanju pa le s priimkom (v bolj neformalnih besedilih pa tudi le z imenom), npr.: »*Darko Krašovec* je bil ponoči, na prvi seji pravkar oblikovane vlade Mira Cerarja, potrjen za generalnega sekretarja. Čeprav do zdaj sodnik, pa **Krašovec** v politiki ni novo ime.« Pri časopisnih naslovih so pogoste tudi katafore take vrste, saj je oseba v naslovu omenjena le s priimkom, v samem članku pa je potem navedena s polnim imenom.

Postopek za to vrsto sklicev pravzaprav ni posebej zapleten, če imamo podatek, kaj so osebna imena, vsa imena oseb je treba shraniti v seznam in potem pogledati po seznamu, kadar naletimo le na posamičen priimek oz. ime.

Zapis, ki ga uporablja Amebisov vmesni jezik, ki prvi del imena osebe (običajno torej osebno ime) zapiše v elementu JED (jedro dela samostalniške fraze), priimek pa v elementu IMP (imenski prilastek), po drugi strani pa sam priimek postane JED (če pa je pred imenom še kakšna druga beseda, npr. »matematik Josip Plemelj«, pa celo tako osebno ime kot priimek postaneta IMP), sicer pomeni, da se postopek malo zaplete in je treba pri izvedbi paziti na vse te pretvorbe. Dodatna težava so primeri, kjer bi morali sklic vezati na element IMP, kar za zdaj še ni podprto (če je torej posamičen priimek uporabljen kot prilastek za drugo besedo, npr. »matematik Plemelj« kot sklic za »matematik Josip Plemelj«).

4.4. Anafore pri osebnih zaimkih

Postopek za razreševanje anafor je bil zasnovan na podlagi metod na osnovi aktivacije (activation-based methods), kakor so opisane v (Němčík, 2006) in ki izhajajo iz dela Eve Hajičove in sodelavcev, vendar v tem trenutku še v precej poenostavljeni in predelani obliki.

Postopek je tak, da se gradi kontekst analize, ki vsebuje seznam kandidatov za razreševanje anafor, pri čemer ima vsak kandidat shranjeno analizo ustrezne samostalniške fraze, mesto zadnje uporabe (npr. osebek, predmet v tožilniku, prislovno določilo), podatke o spolu, številu, osebi in živosti ter oceno. Ko se pride do osebnega zaimka, ki še nima razrešenega sklica, se poišče, ali obstaja kakšen kandidat, ki ustreza glede spola, števila, osebe in živosti, če jih je več, se izbere tisti, ki ima višjo oceno oz. se je pojavil zadnji, dodatno pa oceno zviša še ujemanje mesta uporabe (če npr. razrešujemo sklic pri osebku, ima prednost kandidat, ki je bil že prej osebek).

Uporaba kandidata mu poviša oceno, z začetno oceno se na seznam kandidatov dodajo tudi vse samostalniške fraze, ki nastopajo v analizi. Na koncu vsakega stavka, povedi in odstavka se znižajo (prepolovijo) ocene vseh obstoječih kandidatov, kandidati, katerih ocena pade na 0, se izbrišejo iz konteksta analize.

Ta osnovni postopek je bil dopolnjen z dodatnimi pravili, ki so opisana v nadaljevanju.

4.4.1. Premi govor

Premi govor prekine tok pripovedovanja z drugim tokom, zato konteksta iz spremnega besedila ne smemo uporabiti pri analizi premega govora in obratno. Rešitev je, da ima analizator dva konteksta – enega za osnovno besedilo in drugega za premi govor, pri čemer se kontekst za premi govor vsakič ponastavi (dokler ne bo izdelana boljša analiza diskurza, ki bi določila, kdo se s kom pogovarja).

Dopolnitev za prihodnost je še, da se iz spremnega stavka v kontekst premega govora preneseta prva in druga oseba (iz »Janez je rekel Micki: 'Jutri ti bom prinesel to knjigo.'« bi tako lahko ugotovili, da bo Janez prinesel knjigo Micki).

4.4.2. Pomenske omejitve

Samo informacije o skladnji in osnovne omejitve (oseba, spol, število) ne zadoščajo vedno za razreševanje sklicev.

Metka je prebrala knjigo, ki jo je napisala Karmen, in jo povabila na kavo.
Metka je prebrala knjigo, ki jo je napisala Karmen, in jo vrgla stran.

Slika 3. Pomenske omejitve pri sklicih

Čeprav sta si povedi na sliki 3 enaki do drugega »jo«, je razrešitev tega sklica vseeno drugačna. V prvi povedi se drugi »jo« nanaša na »Karmen«, v drugi pa na »knjiga«.

Podobno je v realnem primeru »Zadremala je že skoro, ali zgodilo se ji je, kakor da bi polagoma drsala navzdol, kakor da bi se skrinja nagibala, nagibala ... in prestrašila se je in se je prebudila.«, kjer se ni prestrašila skrinja, ampak oseba, ki sicer ni navedena v tej povedi.

V precejšnjem delu primerov si bo dalo pomagati že s tem, da imajo glagolske predloge lahko pri parametrih omejitve, ali so ti parametri obvezno osebe (oz. organizacije) oz. niso osebe. Vendar pa to vedno ne zadošča, v primeru »Hm, lahko bi kar takoj pojedel to deklico, ampak je premajhna, da bi mi potešila lakoto. Če

odigram pravilno, bom lahko pojedel njo, pa tudi njeno babico!« je tako postopek najprej menil, da se »njo« nanaša na »lakoto« kar pomeni, da je treba v predlogi omejiti, da se ne da pojesti lakote. V takih primerih bi si lahko pomagali tudi s korpusom, vendar oseb ne jemo prav pogosto, če ne gre za pravljico.

4.4.3. Stavki brez analize

Pojavi se vprašanje, kaj narediti v primeru, ko analizatorju ne uspe analizirati katerega od vmesnih stavkov. Tak primer je bil »Sončni žarki so se že igrali na strehi županove hiše. Francka je bila vsa nemirna, srce ji je utripalo od sreče in obenem od straha, da bi zamudila voz.«, kjer analizator ni prepoznal stavka »Francka je bila vsa nemirna« (ker še ni podpiral kombinacije zaimka »ves« s pridevnikom na mestu povedkovega določila), zato je potem postopek priredil zaimku »ji« vrednost »županova hiša«.

Idealna rešitev je, da se dopolni analizator, vendar ni mogoče pričakovati, da mu bo v dogledni prihodnosti uspelo analizirati vse (še posebej pri izpustih) zato je varianta, ki je vredna razmisleka, ta, da se v takih primerih ponastavi (pobriše) stanje konteksta. Na ta način sicer lahko izgubimo nekatere razrešitve sklicev, ki izhajajo še iz prejšnjih stavkov, vendar se izognemo napakam, kar je v večini primerov bolj pomembno (torej povečamo natančnost na račun priklica).

Vsekakor pa je dolgoročna rešitev izboljševanje analizatorja.

4.4.4. Ponavljanje v stavku

V primeru »Ko sva zapuščala hišo, se je mačka nekako med nogami izmuznila nazaj v hišo. Nisva jo želela pustiti v hiši, ker se neprestano trudi požreti papigo.« je analizator poskušal razrešiti »jo« s »hiša« namesto »mačka«. Pomenske omejitve ni (morala bi biti precej podrobna, kjer lahko nekje pustiš tako osebo kot predmet), možno pa je postaviti dodatno zahtevo, da ne smemo znotraj posameznega stavka razrešiti zaimka z besedo, ki v tem stavku še nastopa, s čimer se potem zaimek »jo« razreši v »mačka«.

4.4.5. Pomenske razlike med izpuščenimi in navedenimi osebnimi zaimki

Lastnost, ki nam lahko pomaga pri razreševanju sklicev, je, da se na neživo vedno nanašamo le z naslonsko obliko osebnega zaimka. Tako ne moremo reči »Knjiga je bila zelo zanimiva in njo sem prebral v dveh urah.« ampak le »Knjiga je bila zelo zanimiva in prebral sem jo v dveh urah.« Ker v imenovalniku ni naslonskih oblik, to pomeni, da neživo v osebku ne more biti nadomeščeno z osebnim zaimkom, ampak le z izpustom osebnega zaimka ali pa s kazalnim zaimkom.

S pomočjo tega pravila lahko v besedilu »Njegova jeza je v Naomi zbujala občutek krivde. Ona je bila tista, ki je vztrajala na prenovi strehe.« ugotovimo, da »Ona« ne smemo razrešiti z »njegova jeza«, ampak z »Naomi«.

4.5. Prislovni zaimki

Tukaj se razrešujejo sklici, ki so vezane ne prislovne zaimke, in sicer na »tam«, »tja« in »takrat«. Prva dva se sklicujeta na kraj, drugi pa na čas.

Oblaki nastajajo poleti *nad večjimi ognjeniki*. **Tam** nastanejo zato, ker se topli zrak dviga in ohlaja.
Princ Borjatinski, guberner Jakutska je leta 1670 zaupal Dežnjovu odpravo *v Moskvo*. **Tja** je moral odnesti »sobolji zaklad« in uradne dokumente.
Biologija se je začela hitro razvijati in rasti, *ko je Anton van Leeuwenhoek izboljšal mikroskop*. **Takrat** so učenjaki odkrili semenčice, bakterije, infuzorije in raznovrstnost mikroskopskega življenja.
Večina dragocenosti, ki jo jih Slovani naropali v Hersonu, se je znašla v Novgorodu, kjer so jih vse do 20. stoletja hranili *v katedrali sv. Sofije*. **Tja** so prišle morda po zaslugi prvega novgorodskega škofa Joahima Hersonskega, katerega ime kaže na njegovo povezanost s tem mestom.
Prosti čas je mojster izkoristil za obisk *Londona*. **Tja** je prišel kot izrazit skladatelj italijanske opere.

Slika 4. Primeri s prislovnimi zaimki

Slika 4 vsebuje primere prislovnih zaimkov s sklici. V prvih treh primerih zaimek nadomešča prislovno določilo iste vrste (z isto vprašalnico), četrti primer pa kaže, da je pri krajevnih prislovnih določilih treba imeti možnost, da se pretvarja med prislovnimi določili kraja za »kje« in »kam« (torej je treba »v katedrali sv. Sofije« pretvoriti v »v katedralo sv. Sofije«), kar za zdaj v Amebisovem vmesnem jeziku ni možno, zato bo treba ustrezno dopolniti podatkovno bazo Ases s povezavami med pomeni predlogov oz. prislovov.

Zadnji, peti primer pa kaže, da ni nujno, da se prislovni zaimki sklicujejo na prislovna določila, ampak se lahko sklicujejo tudi na samostalnike, npr. na zemljepisna imena.

Razmislek pri prislovnih zaimkih je, da so relativno redki v besedilih, uporabljamo jih le, če želimo povezavo posebej poudariti. Veliko pogostejše je implicitno navezovanje, da se naslednji stavek dogaja v istem času in prostoru, zato bo treba razmišljati tudi v smeri, kako najti te implicitne sklice.

4.6. Katafora v osebku odvisnika

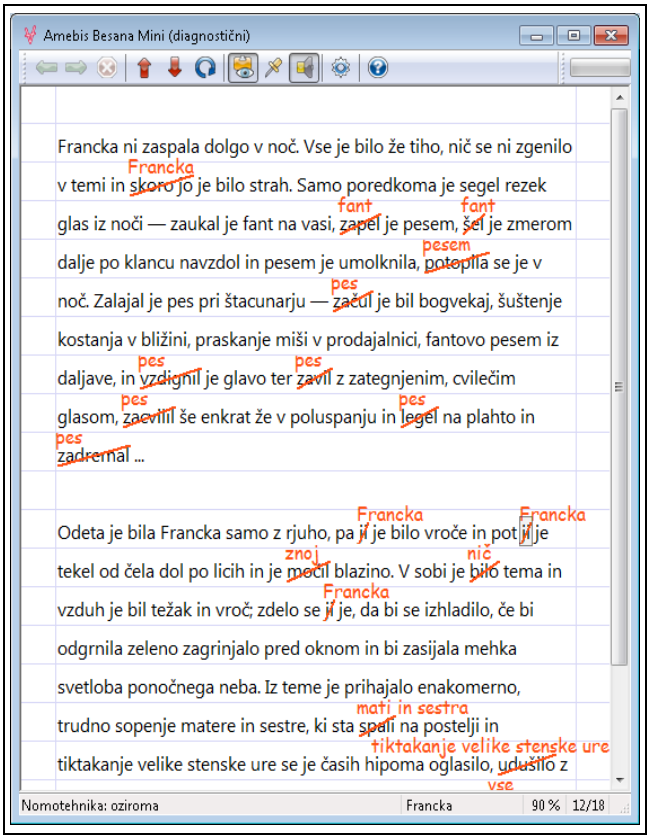
Pri katafori je zaimek pred samostalniško frazo, ki jo nadomešča. Primer za to je poved »Ker jo je zeblo, je Mojca oblekla jopico.«

Vidimo lahko, da je pri tem tipu zaimek, ki ga želimo razrešiti, v odvisniku, razrešitev sklica moramo pa poiskati v osebku glavnega stavka, ki sledi, pri čemer pa moramo paziti še na to, da je ta osebek na začetku stavka, drugače imamo težave pri primeru »Ker jo je zeblo, ji je Mojca oblekla jopico.« ali pa »Ker jo je zeblo, ju je Mojca poslala domov.«

5. Rezultati

Za lažje preizkušanje delovanja razreševanja sklicev (neposredno branje Amebisovega vmesnega jezika ni posebej preprosto, saj je bolj prilagojen temu, da ga berejo računalnik) je bila zgrajena posebna verzija slovničnega pregledovalnika Besana, in sicer z vmesnikom Besana Mini. Sklici se izpisujejo kot ena od napak, ki jih program išče (razrešitve (izpisane vedno v imenovalniku) nadomestijo osebni zaimek, če pa gre za izpust osebnega zaimka, pa nadomestijo glagol), za boljšo preglednost se izključi izpis vseh drugih napak. Na ta način se besedilo, s

katerim želimo preveriti delovanje razreševanje sklicev, le skopira v odložišče in Besana takoj izpiše najdene razrešitve, kot je primer na sliki 5.



Slika 5. Primer razreševanja sklicev na prvih dveh odstavkih romana Na klancu

Prvi rezultati (kot na primer zgornji primer) so videti obetavno, kaže pa se, da se bo treba bolj posvetiti izboljšavam analizatorja (zdaj denimo v primeru »Jetra so za vretenčarje značilen organ. Imajo osrednjo vlogo v presnovi in številne druge naloge« tako v drugi povedi ne najde izpusta osebnega zaimka, ker analizator napačno določi, da je osebek »naloge«).

Težave nastanejo tudi zato, ker sistem še ne vsebuje dovolj pomenskih omejitev, ki pomagajo pri izbiranju pravega sklica. Te omejitve bi tudi v splošnem pomagale pri razdvoumljanju, zato bo dopolnjevanje baze Ases v tej smeri zelo koristno.

Občasno se pokažejo tudi težave, ker starejši kontekst preveč vpliva na nove stavke, kar kaže, da bo smiselno preizkusiti hitrejše pozabljanje konteksta. Točne nastavitve teh parametrov pa bodo zahtevale več preizkušanja in predvsem pripravo korpusa primerov razrešenih sklicev, kar bo omogočilo hitrejše preizkušanje različic.

6. Uporaba v sistemu Piflar

Piflar je sistem za odgovarjanje na vprašanja v naravnem jeziku, ki se med drugim uporablja na Amebisovem portalu za virtualne asistente SecondEgo (http://www.secondego.com). Sistem kot osnovo uporablja Amebisov vmesni jezik, velika omejitev pa so bili zaimki v vhodnem besedilu, ker se je sistem naučil znanje z zaimki namesto z njihovimi pravimi pomeni (Holozan, 2014)

Že v (Vicedo, Ferrández, 2000) je bilo pokazano, da je razreševanje sklicev pomembno za odgovarjanje na vprašanja. Zato je bil Piflar dopolnjen s podporo za element ORI, ki je bil dodan v vmesni jezik za zapisovanje razrešenih sklicev, tako da zdaj uporablja ta element namesto originalnega jedra (JED). S to dopolnitvijo zdaj pravilno odgovarja tudi v primerih, ko je treba upoštevati sklice, kar prikazuje tabela 1, kjer so odebeljeno označena vprašanja, na katera je mogoče odgovoriti zaradi razrešenih sklicev, prej pa bi bili uporabljeni osebni zaimki oz. Piflar ni imel odgovora na vprašanje.

vprašanje	kratki odgovor	dolgi odgovor	prejšnji kratki odgovor
Ali je Miha prebral knjigo?	da	Da.	da
Kdo je prebral knjigo?	Miha	Knjigo je prebral Miha.	Miha
Kaj je Miha prebral?	knjigo	Miha je prebral knjigo.	knjigo
Ali je Miha potem pojedel kosilo?	da	Da.	/
Kdaj kosilo je pojedel Miha?	potem	Kosilo je pojedel Miha potem.	/
Kaj je Miha pojedel potem?	kosilo	Miha je pojedel kosilo potem.	/
Kdo je pojedel kosilo potem?	Miha	Kosilo je pojedel Miha potem.	on
Ali je Miha šel v Ljubljano?	da	Da.	/
Kdo je šel v Ljubljano?	Miha	V Ljubljano je šel Miha.	on
Kam je šel Miha?	v Ljubljano	Miha je šel v Ljubljano.	/
Ali je Miha naletel na Janeza?	da	Da.	/
Na koga je Miha naletel?	Janez	Miha je naletel na Janeza.	/
Kdo na Janeza je naletel?	Miha	Na Janeza je naletel Miha.	on
Ali je Miha pozdravil Janeza?	da	Da.	/
Kdo je pozdravil Janeza?	Miha	Janeza je pozdravil Miha.	on
Koga je Miha pozdravil?	Janeza	Miha je pozdravil Janeza.	/

Tabela 1: Seznam vprašanj in odgovorov, ki jih najde Piflar za primer »Miha je prebral knjigo. Potem je pojedel kosilo in šel v Ljubljano. Srečal je Janeza in ga pozdravil.«.

Podoben primer, ki pa vsebuje še prislovni zaimek, je, če imamo vhodno besedilo »Matematik Josip Plemelj se je rodil 11. decembra 1873 na Bledu. Tam je obiskoval osnovno šolo.« Na vprašanje »Kje je Josip Plemelj obiskoval osnovno šolo?« tako dobimo odgovor »Josip Plemelj je hodil v osnovno šolo na Bledu.« Težava pri teh krajevnih (in podobno časovnih) določitvah pa je, da so največkrat implicitne (se prenašajo iz prejšnjih stavkov brez izrecne uporabe prislovnih zaimkov), česar Piflar še ne zna uporabiti.

Sistem Piflar je bil dodatno dopolnjen s tem, da zna odgovarjati tudi na vprašanja, ki niso zastavljena v obliki stavka, ampak le kot posamičen stavčni člen. Tako npr. kot odziv na vprašanje »Miha« poišče dejstvo, ki vsebuje samostalniško frazo »Miha«, npr. »Miha je šel v Ljubljano.«, če ima v bazi primer iz Tabele 1. Na ta način se Piflar bolj uspešno odziva na način iskanja, na katerega so uporabniki navajeni iz običajnih iskalnikov.

Pri tovrstnih vprašanjih je možno kombinirati tudi npr. osebek in prislovno določilo. Če imamo npr. učni stavek »Isaac Newton je umrl 20. marca 1727 v Kensingtonu.«, zdaj Piflar na vprašanje »Isaac Newton 1727« odgovori: »Isaac Newton je umrl 20. marca 1727.«

Pri uporabi v sistemu Piflar pa se z bolj kompleksnimi odgovori kaže tudi to, da bo treba dopolniti tudi generator, ki prevaja vmesni jezik v naravni jezik, in sicer v smeri, da bo poskrbel za naravnejše odgovore s tem, da bo dodajal izpuste in po potrebi tudi sklice z osebnimi zaimki, da bodo odgovori zveneli bolj naravno. Zdaj npr. pri učnem besedilu »Oblaki nastajajo poleti nad večjimi ognjeniki. Tam nastanejo zato, ker se topli zrak dviga in ohlaja.« na vprašanje »Zakaj nastanejo oblaki nad večjimi ognjeniki?« odgovori »Do oblakov pride nad večjimi ognjeniki, ker se dviguje topli zrak in ker se ohlaja.«, namesto »Do oblakov nad večjimi ognjeniki pride, ker se topli zrak dviguje in ohlaja.«

7. Sklep

Razreševanje sklicev se je pokazalo kot uporabno tako v sistemu Piflar kot tudi v strojnem prevajalniku Presis (ki tako zdaj poved »Pobral sem knjigo in jo začel brati.« prevede v »I picked up a book and started to read it.« namesto v »I picked up a book and started to read her.« kot do zdaj).

Potencialna možnost uporabe je še pri iskanju po korpusih, npr. pri iskanju kolokacij, kjer bi z razširitvijo iskanja na osebne zaimke z razrešenimi sklici lahko povečali število zadetkov pri isti velikosti korpusa.

Sistemu še ne uspe razrešiti vseh sklicev, zato je še veliko možnosti za izboljšave, še posebej to velja za sklice, ki niso zaimki, niti dotaknil pa se še ni tudi bolj zapletenih povezanih sklicev (par – moški).

Razreševanje sklicev je možno izboljšati tudi z analizo diskurza, predvsem dialogov, s čimer bi se lahko bolje povezale informacije v različnih odstavkih in v premem govoru. Tak primer je npr. v sliki 6.

Rdeča kapica je vprašala volka: »Zakaj imaš tako veliko oči?« »Da te bolje vidim.«

Slika 6. Primer diskurza.

Na podlagi tega učnega besedila, bi moral biti Piflar sposoben na vprašanje »Zakaj ima volk tako velike oči?« odgovoriti z »Volk ima tako velike oči, da bolje vidi Rdečo kapico.«

Pogoj za nadaljnji razvoj pa bo verjetno tudi priprava korpusa primerov razrešenih sklicev, ki bi omogočil hitro primerjavo delovanja različnih postopkov, pri pripravi takega korpusa pa bi bilo pomembno, da se ne omeji le na osebne zaimke, ampak se označijo tudi druge vrste sklicev, da bo tak korpus uporaben tudi za preizkušanje razreševanja bolj zapletenih vrst sklicev.

8. Literatura

Brennan, S. E., Friedman M. W., Pollard C. J., 1987. A centering approach to pronouns. V *Proceedings of the 25th Annual Meeting of the ACL*. Stanford. 155–162.

Erjavec, T., Krek, S., 2008: Oblikoskladenjske specifikacije in označeni korpusi JOS, V T. Erjavec, J. Žganec Gros (ur.), *Zbornik 6. konference Jezikovne tehnologije 2008*. Ljubljana: IJS.

Ge, N., Hale, J., Charniak E., 1998. A statistical approach to anaphora resolution. V *Proceedings of the Sixth Workshop on Very Large Corpora*.

Hobbs, J. R. 1978. Resolving pronoun references. V Barbara J. Grosz, Karen Spärck-Jones, and Bonnie Lynn Webber (ur.), *Readings in Natural Language Processing*. Morgan Kaufmann Publishers, Los Altos. 339–352.

Holozan, P., 2011. *Samodejno izdelovanje besedilnih logičnih nalog v slovenščini.* Magistrsko delo, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko.

Holozan, P., 2014. Piflar: sistem za učenje in odgovarjanje na vprašanja v naravnem jeziku. V M. Orel in S. Jurjevčič (ur.), *MEDNARODNA konferenca InfoKomTeh*. Polhov Gradec: Eduvision.

Joshi, A. K., Kuhn, S., 1979. Centered logic: The role of entity centered sentence representation in natural language inferencing. V *Proceedings of the International Joint Conference on Artificial Intelligence*, Tokyo. 435–439.

Lappin, S., Leass. H. J., 1994. An algorithm for pronominal anaphora resolution. *Computatinal Linguistics*, 20(4). 535–561.

McShane, M. Beale, S., Nirenburg, S., 2010. Reference Resolution Supporting Lexical Disambiguation. V zborniku *2010 IEEE Fourth International Conference on Semantic Computing*. Los Alamitos: IEEE Computer Society.

Mitkov R., 1999. *Anaphora Resolution: The State Of The Art.* Working paper, University of Wolverhampton.

Mitkov, R., Evans C., Orasan, C., 2002. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. V *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Mexico City, Mexico, February, 17 – 23.

Němčík, V., 2006. *Anaphora Resolution.* Magistrsko delo, Masarykova universita, Fakulta informatiky.

Vicedo, J. L., Ferrández, A., 2000. Importance of Pronominal Anaphora resolution in Question Answering systems. V *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL}*. 555–562.