UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Tadej Štajner

Razločevanje entitet v besedilih s strojnim učenjem in predznanjem

DIPLOMSKO DELO
NA UNIVERZITETNEM ŠTUDIJU

Mentor: doc. dr. Janez Demšar

Somentorica: doc. dr. Dunja Mladenić

Ljubljana, 2009

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Namesto te strani **vstavite** original izdane teme diplomskega dela s podpisom mentorja in dekana ter žigom fakultete, ki ga diplomant dvigne v študentskem referatu, preden odda izdelek v vezavo!

IZJAVA O AVTORSTVU

diplomskega dela

Spodaj podpisani Tadej Štajner,

z vpisno številko 63030061,

sem avtor diplomskega dela z naslovom:

Razločevanje entitet v besedilih s strojnim učenjem in predznanjem

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom
  doc. dr. Janeza Demšarja
  in somentorstvom
  doc. dr. Dunje Mladenić
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 9.6.2009 Podpis avtorja:

Zahvala

Za pomoč in podporo pri izdelavi te diplomske naloge se zahvaljujem mentorju doc. dr. Janezu Demšarju in somentorici doc. dr. Dunji Mladenić. Prav tako bi se rad zahvalil tudi sodelavcem Odseka za tehnologije znanja Inštituta Jožef Stefan, še posebej Marku Grobelniku, Blažu Fortuni in Mitju Trampušu za dobre ideje in spodbudno delovno okolje. Nenazadnje bi se za podporo zahvalil tudi moji družini in punci Alenki.

Kazalo

Seznam uporabljenih kratic in simbolov

URI : *Uniform Resource Identifier*, enotni označevalnik vira. Je niz podatkov, ki enolično določa nek vir na omrežju, dosegljiv z določenim protokolom.

RDF : *Resource Description Framework*, jezik za opis razmerij med viri glede na vnaprej definirano ontologijo s poudarkom na ponovni uporabi ontologije.

SI : *Specific Pointwise Mutual Information*, specifična skupna informacija dveh spremenljivk. Kvantificira odstopanje med verjetnostjo sopojavitve spremenljivk glede na hkratno pojavitev in sopojavitvijo glede na verjetnost neodvisne pojavitve obeh spremenljivk.

IPS : *Instance Participation Selectivity*, mera ki ocenjuje koliko je relacija med entitetama določenih tipov selektivna glede na pogostost tovrstne relacija. Bolj kot je relacija selektivna, višjo težo ima pri razločevanju.

Povzetek

Metode strojnega učenja se uspešno uporabljajo pri analizi besedil. Ker pa je naravni jezik dvoumen, je potrebno določiti, katero entiteto predstavlja posamezna pojavna oblika. S tem problemom se ukvarjajo metode razločevanja entitet.

V nalogi smo preučili, kako doseči čim višjo natančnost razločevanja entitet z uporabo statističnega učenja in predznanja. Primerjali smo natančnost razločevanja posamičnih entitet z zmogljivostjo skupinskega razločevanja. Preučili smo tudi vpliv uporabe predznanja. V ta namen smo definirali metodo, na podlagi katere je moč v skupinskem razločevanju uporabiti tako implicitne, kot tudi eksplicitne relacije med entitetami. Ugotovili smo koristnost uporabe vsebinske podobnosti in statistično naučenih sopojavitev kot možnosti izražanja implicitnih relacij. Preučili smo tudi možnost uporabe eksplicitnega relacijskega predznanja za skupinsko razločevanje, tako da smo določili pomembnost posameznih tipov relacij.

Ključne besede:

razločevanje entitet, ontologija, predznanje, statistično učenje, obdelava naravnega jezika, tekstovno rudarjenje

Abstract

Machine learning methods are being successfully applied in text mining. Because of ambiguities which are inherently present in natural languages, we are faced with a challenge of determining the actual identities of entities mentioned in a document. Disambiguation is a problem that can be successfully solved by entity resolution methods.

This thesis studies various possibilities for improving entity resolution performance by using various types of background knowledge and statistical learning. We compare precision and recall of pair-wise entity resolution with collective resolution. We also study the possibility of employing background knowledge. For this purpose, we define a multi-relational entity resolution approach, capable of representing implicit as well as explicit relationships. We discover the benefits of using entity co-occurrences and content similarities as implicit relationships. We also propose an approach capable of handling such heterogeneous relations for collective entity resolution.

Key words:

entity resolution, ontology, background knowledge, statistical learning, natural language processing, text mining

Poglavje 1

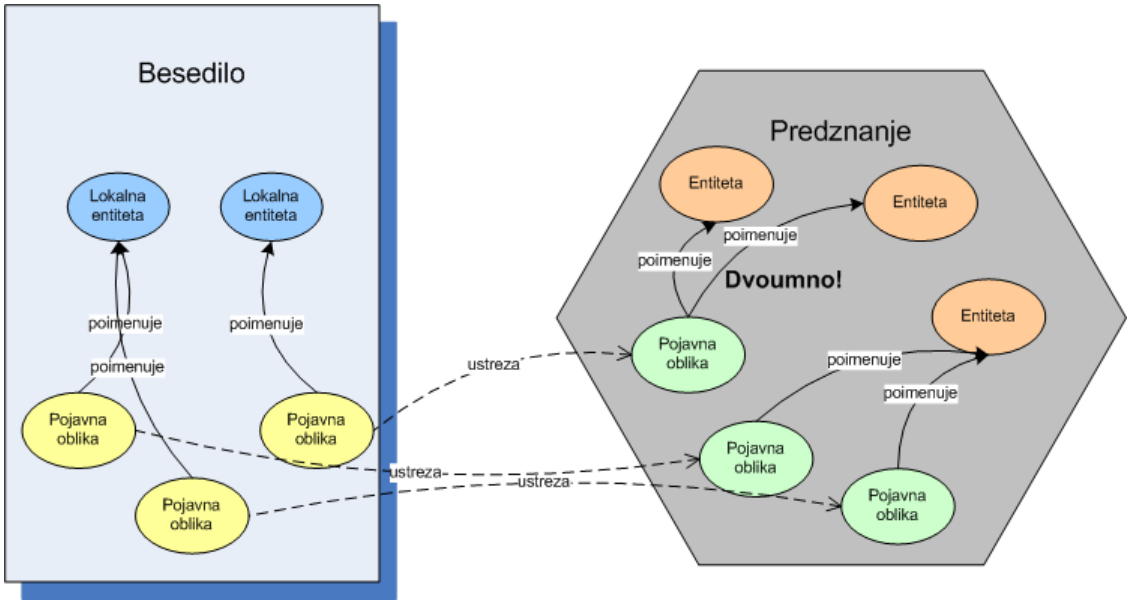
Uvod

1.1 Motivacija

Integracija in deljenje podatkov preko različnih virov informacij je osnova za inteligenten in učinkovit dostop do heterogenih virov informacij. Pogost primer takšnega problema je povezovanje besedila s strukturiranimi in delno strukturiranimi podatkovnimi viri, saj je pogosto veliko znanja predstavljenega z besedilom in ne v bolj eksplicitni obliki. Za premoščanje te razlike je potrebna identifikacija entitet v prostem besedilu, ki je postopek ekstrakcije delno strukturiranih podatkov. Ker pa je jezik besedila dvoumen in je zaradi tega določanje točnega pomena besedne zveze zahteven problem, se poslužujemo različnih modelov, hevristik in tehnik, ki nam izboljšajo natančnost rezultata te odločitve. Določitev pomena neke besedne zveze je torej bistvena za tovrstno semantično integracijo. Želimo tudi ugotoviti, kako lahko različne tipe predznanja izkoristimo za razločevanje entitet, tako da bo metoda uporabna na čim širšem razponu podatkovnih virov.

1.2 Problemska domena

Metode strojnega učenja se uspešno uporabljajo pri analizi besedil [1]. Izziv pri analizi že vrsto let predstavlja naravni jezik s svojo dvoumnostjo, saj so enake besede pogostokrat uporabljene v različnih kontekstih z različnimi pomeni. Temeljni problem, s katerim se srečujemo, je pravilna odločitev pri določanju pomena neke pojavne oblike. V predstavljeni nalogi je pomen enolično določen z entiteto iz predznanja, naš cilj pa je ugotoviti, kateri entiteti iz predznanja ustreza določena pojavna oblika iz dokumenta.



Slika 1.1: Razločevanje entitet v besedilu

Da lahko govorimo o pomenu neke besede, je koristno imeti tudi predznanje, ki vsebuje podatke tako o možnih pomenih neke pojavne oblike, kot tudi o možnih poimenovanjih nekega koncepta. S tem predznanjem nato pri problemu dvoumnosti definiramo prostor rešitev problema. Nato je naš izziv najti v predznanju najprimernejšo entiteto, ki ustreza tisti v besedilu. Prikaz te situacije vidimo v 1.1, kjer je tudi ponazorjen problem dvoumnosti, ki ga moramo razrešiti.

S problemom določanja pomena neke pojavne oblike v besedilu se ukvarjajo metode razločevanja entitet. Problemi, analogni razločevanju entitet, se pojavljajo na različnih področjih. Teoretični temelji so definirani v teoriji povezovanja zapisov [2], sorodni izzivi pa so integracija podatkov iz različnih podatkovnih baz [3, 4], identifikacija objektov [5], odpravljanje duplikatov v podatkovnih bazah [6] in ugotavljanje pomena pojavne oblike besede [7, 8].

V splošnem je razločevanje entitet postopek, ki v grafu entitet združi tista vozlišča, ki ustrezajo kriteriju podobnosti, kot je na primer podobnost med besedilom dokumenta in opisom entitete iz predznanja, lahko pa podobnost zajema tudi relacijsko podobnost [9]. Reševanje problema dvoumnosti pojavnih oblik je specializacija razločevanja entitet, kjer je prostor možnih razločevanj definiran vnaprej. Potencialni prostor kandidatov iz predznanja za neko entiteto iz dokumenta predstavljajo vse entitete iz predznanja, ki so poimenovane

s pojavno obliko entitete iz dokumenta. Ker predpostavljamo, da so entitete iz predznanja konsistentne in brez duplikatov, jih v našem postopku ne združujemo.

Metode za razločevanje entitet lahko delimo po več kriterijih:

- Glede na vir dodatnih podatkov za razločevanje:

  Razločevanje na osnovi sopojavitev parov entitet v besedilu je uporabno v situacijah, ko imamo na voljo že označena besedila, iz katerih se lahko naučimo statistiko sopojavitev.

  Razločevanje s pomočjo atributnega predznanja vsebuje entitete in podatke o njih uporabno v situacijah, ko imamo na voljo opisne podatke o entitetah, kar je tipično za opisne enciklopedijske podatkovne vire.

  Razločevanje s pomočjo relacijskega predznanja vsebuje entitete in njihove medsebojne povezave uporabimo v situacijah, ko imamo na voljo graf entitet z medsebojnimi povezavami, tipično za socialna omrežja, taksonomije ali semantične enciklopedijske podatkovne vire.

  V tej nalogi tudi definiramo model, na katerem prevedemo te tri metode v skupno domeno, na kateri potem izvajamo samo razločevanje entitet brez posebnega ozira na sam vir podatkov. Na tem nivoju pa poznamo naslednje tehnike:

- Posamično ali skupinsko razločevanje:

  Posamično razločevanje je definirano kot problem označevanja z najprimernejšo entiteto iz predznanja tako, da je ta odločitev neodvisna od drugih razločevanj v besedilu.

  Skupinsko razločevanje uporablja tudi soseščino med entitetami in ne predpostavlja neodvisnosti od predhodnih razločevanj v istem besedilu.

Prispevek te naloge je, da definira splošno ogrodje za skupinsko razločevanje entitet s pomočjo medsebojne povezanosti entitet, kjer o sami naravi povezanosti ne predpostavimo ničesar. To nam omogoča, da si to povezanost razlagamo bodisi kot eksplicitne relacije, sopojavitve ali podobnost, kar nam daje fleksibilnost pri uporabi. Ta naloga v nadaljnjih poglavjih za vsak posamezen tip vira podatkov določi, kako definirati soseščino entitet, ki jo potrebujemo za skupinsko razločevanje.

1.3 Struktura diplomske naloge

Opisan problem leži med področji razločevanja entitet, statističnega učenja in semantičnega spleta. V prejšnjem poglavju smo definirali problemsko domeno, v nadaljevanju pa predstavimo obstoječe pristope in jih umestimo glede na vir in način uporabe predznanja.

Nato sledi opis uporabljenih podatkovnih virov, tako za semantično predznanje, kot tudi za korpus besedila, v katerem želimo identificirati entitete. Opišemo, kako lahko uporabimo opise in atribute entitet, relacije med njimi in njihove sopojavitve v korpusu.

V nadaljnjih poglavjih definiramo, kako te različne pristope transparentno prevedemo na skupen postopek razločevanja entitet in jih nato obravnavamo na enak način. Definiramo razločevanje na podlagi atributne podobnosti, razločevanje na podlagi medsebojnih relacij in razločevanje na podlagi statistike sopojavitev. Predlagamo tudi način združevanja rezultatov vseh pristopov.

Na koncu obravnavane pristope tudi ovrednotimo z eksperimentom, v zaključku pa podamo naše vrednotenje pristopov in ugotovitve.

Poglavje 2

Pregled obstoječih pristopov

Iz stališča procesiranja naravnega jezika naš pristop spada med sisteme za določanje pomena pojavne oblike s pomočjo predznanja. Medtem ko [10] predlaga nenadzorovano določanje dvoumnosti pojavnih oblik, se predstavljena naloga poslužuje predznanja, ki opisuje entitete in njihove možne pojavne oblike, s čimer si definiramo prostor problema, ki ga rešujemo.

2.1 Predznanje o entitetah

V literaturi pogosto najdemo uporabo dodatnih podatkov za izboljšavo natančnosti razločevanja entitet iz besedila [11, 12].

Za namene te naloge je predznanje izraženo kot ontologija - graf entitet, ki so opisane z atributi in med seboj povezane z različnimi relacijami. Tak model je moč sestaviti iz podatkov, izraženih v obliki RDF [13] (*Resource Description Framework*). Ta oblika predpisuje podatke v obliki trojic ⟨objekt, relacija, subjekt⟩, kjer so posamezni elementi trojice predstavljen z enotnimi označevalniki vira (Uniform Resource Identifier - URI), subjekt pa je lahko tudi dobesedna vrednost. Navedena ontologija je dovolj splošna, da nanj lahko prevedemo tudi entitetno-relacijske ali razredne modele [14].

Ker pa reševanje zastavljenega cilja poteka na nivoju entitet in ne le trojic, dano množico trojic prevedemo na graf, tako da vsak enotni označevalnik vira predstavlja svojo entiteto, predikati pa so povezave med entitetami. Poleg tega želimo tudi vedeti, pod katerimi pojavnimi oblikami se lahko pojavi neka beseda, zato je koristno poleg enotnega označevalnika poznati tudi oznake, ki predstavljajo to entiteto, četudi niso nujno enolične.

Predznanje KB formalno predstavimo kot množico trojic, sestavljenih iz entitet $e \in E$, ki so v relaciji $p \in R$ z drugimi entitetami ali atributi $a \in A$.

$$KB = \{\langle s, p, o\rangle; s \in E \wedge o \in (E \cup A) \wedge p \in R\} \tag{2.1}$$

Entitete vsebujejo tudi različne vrste atributov - posebej obravnavamo pojavne oblike in besedilo, ki opisuje entiteto. Atributi, ki izražajo pojavne oblike entitete, so v relaciji rdfs:label[1] z entiteto, medtem ko je opis niz, ki je v relaciji rdfs:comment[2] z entiteto.

$$\text{PojavneOblike}(e_i) = \{a \in A; p \in R \wedge p = \text{rdfs:label} \wedge \langle e_i, p, a\rangle \in KB\} \tag{2.2}$$

$$\text{Opis}(e_i) = \{a \in A; p \in R \wedge p = \text{rdfs:comment} \wedge \langle e_i, p, a\rangle \in KB\} \tag{2.3}$$

- Definiranje prostora možnih rešitev (podatki o entitetah in njihovih pojavnih oblikah).
- Atributi entitet, ki jih uporabimo za določanje medsebojne vsebinske podobnosti. Vsebujejo preproste lastnosti entitet, med drugim tudi besedni opis.
- Eksplicitne relacije med entitetami, ki predstavljajo soseščine entitet.

Najpogostejša je uporaba opisnega predznanja entitet, ki v splošnem zajema definiranje slovarja entitet in njihovih atributov, kot jih najdemo v [11, 15, 16]. V [12] je avtor uporabil tudi kategorije iz Wikipedije tako za definiranje prostora rešitev, kot tudi za ocenjevanje ustreznosti posameznega pomena na podlagi podobnosti v modelu vektorskega prostora besedila [17].

Nekateri pristopi tudi izkoristijo medsebojne povezave med entitetami, ki lahko tudi vsebujejo pomembne informacije. [18] uporablja PageRank [19] na grafu relacij med entitetami za ocenjevanje relevantnosti pomena. Primer uporabe statističnega relacijskega učenja za razločevanje entitet lahko vidimo tudi v [20]. Ker imajo različni tipi relacij med entitetami lahko različno težo, [21]

[1]Relacijo identificira URI http://www.w3.org/2000/01/rdf-schema#label, rdfs je okrajšava imenskega prostora RDF schema, ki določa to relacijo

[2]Relacijo identificira URI http://www.w3.org/2000/01/rdf-schema#comment

predlaga prilagodljivo metodo za določanje pomembnosti relacij, [22] pa definira več hevristik, glede na katere lahko določamo pomembnost posameznega tipa semantične relacije. V nadaljevanju te diplome eno izmed predlaganih metod za določanje pomembnost tudi obravnavamo in ovrednotimo za namen skupinskega razločevanja.

[23] opaža, da imajo entitete, ki nastopajo v istem dokumentu, relativno visoko medsebojno relacijsko povezanost, kar daje medsebojnim relacijam uporabno vrednost pri določanju pomena.

2.2 Uporaba statističnega predznanja

Poleg statičnega atributnega in relacijskega znanja pa lahko natančnost razločevanja izboljšamo tudi z uporabo korpusa besedila, kjer so entitete že identificirane. Te podatke lahko nato izkoristimo za učenje statističnega modela. Splošen model verjetnostnega določanja pomena s pomočjo relacijskega znanja definira [8], [24] pa predlaga generativni model razločanja entitet na osnovi sopojavitev parov entitet.

Sopojavitve dogodkov so pogost vir učnih podatkov za reševanje problemov, sorodnih razločevanju entitet. Primera uporabe te tehnike za ugotavljanje pravilne entitete na podlagi pogostih sopojavitev sta [25], ki jo uporabi za identifikacijo imen proteinov, in [26], ki na ta način razločuje geografske lokacije. [27] sopojavitve uspešno uporablja za nenadzorovano ugotavljanje sinonimov besed, kar je problem, soroden razločevanju entitet.

2.3 Posamično in skupinsko razločevanje

Problem identifikacije in določanja pomena lahko formuliramo tudi kot problem razločevanja entitet. Izvirni matematični model, kot ga definirata Fellegi in Sunter [2], definira karakteristike entitet, na podlagi katerih se določi verjetnost, da sta dve entiteti enaki. Ker predpostavlja neodvisnost med posameznimi odločitvami, je to primer posamičnega razločevanja.

Vprašanje razločevanja entitet se pogosto pojavlja v kontekstu integracije podatkovnih baz, ko imamo opravka z združevanjem različnih množic podatkov brez skupnega ključa, ki bi enolično določal enakost med entitetami iz različnih virov. [28, 29] definirajo razliko med posamičnim in skupinskim razločevanjem in obravnavajo strukturirane podatke, medtem ko v tej diplomski nalogi ta izziv prevedemo na domeno integracije nestrukturiranih in delno strukturiranih podatkov. V domeni razločevanja entitet v besedilu je ena od entitet vedno

lokalna entiteta posameznega dokumenta, ki se primerja z možnimi kandidati iz predznanja.

Pri skupinskem razločevanju lahko entitete med seboj primerjamo z dvema meriloma podobnosti:

atributna podobnost, kjer se primerja vsebino atributov entitet in

aelacijska podobnost, ki pa primerja skupno soseščino entitet.

V nalogi bomo s poskusi na realnih podatkih ovrednotili naslednje pristope:

- posamično razločevanje z vsebinsko primerjavo,
- skupinsko razločevanje z vsebinsko primerjavo,
- skupinsko razločevanje z uporabo heterogenih relacij,
- skupinsko razločevanje z uporabo sopojavitev entitet.

V nalogi preučimo, kako doseči čim višjo natančnost razločevanja entitet z uporabo strojnega učenja in predznanja. S poskusi bomo ugotovili, kako lahko dodaten podatek o sopojavitvi entitet vpliva na natančnost razločevanja entitet. Preučili smo tudi vpliv uporabe predznanja in primerjali natančnost razločevanja posamičnih entitet z zmogljivostjo skupinskega razločevanja. Preučili smo možnost uporabe relacijskega predznanja za skupinsko razločevanje, tako da smo s pomočjo strojnega učenja določili pomembnost posamezne relacije.

Poglavje 3

Priprava podatkov

Postopek je zastavljen tako, da potrebuje za delovanje dva vira podatkov:

- korpus besedila, v katerega želimo identificirati entitete;
- zbirka entitet, ki predstavlja naše predznanje.

Podrobnejši ogled arhitekture postopka je prikazan v dodatku 1.

3.1 Korpus besedila

Besedilo, v katerem identificiramo entitete za potrebe te naloge je korpus časopisa New York Times [30], ki obsega 1.8 milijona člankov, objavljenih v New York Times med leti 1987 in 2007. Poleg samega besedila so članki tudi opremljeni z dodatnimi metapodatki, kot so vsebinske kategorizacije, oznake in identificirane entitete. Ker so entitete normalizirane, smo si z njimi tudi pomagali pri preverjanju natančnosti naše metode.

3.2 Predznanje

V splošnem podatkovni model predvideva, da imajo entitete v predznanju:

- svoje enolične identifikatorje,
- nabor pojavnih oblik, ki jih lahko predstavljajo,
- opis,
- druge atribute,

- relacije z drugimi entitetami.

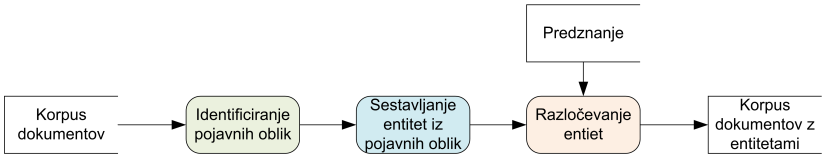
Medtem ko sta prvi dve točki za delovanje nujni, so ostale zaželene. Nekatere izmed metod namreč potrebujejo dodatne podatke. Opis in atributi entitete se uporabljajo pri razločevanju z vsebinsko primerjavo, relacije z drugimi entitetami pa pri določanju uteži relacij v primeru skupinskega razločevanja s pomočjo teh uteži.

Kot predznanje, ki tem zahtevam ustreza, smo izbrali podatkovno zbirko DBpedia [31]. Izvira iz proste enciklopedije Wikipedia, zato ima širok nabor entitet, njihovih pojavnih oblik in atributov. Ker pa želimo obravnavati tudi podatke, ki so relacijsko bolj povezani, smo poleg tega uporabili tudi YAGO [32], ontologijo, ki vsebuje mnogo več relacij med entitetami, kar potrebujemo za določanje njihove pomembnosti. Med obema ontologijama na srečo obstaja neposredna preslikava entitet, zato smo jih združili v eno skupno predznanje, ki ga kot takega v naslednjih poglavjih obravnavamo.

Poglavje 4

Razločevanje entitet iz besedila

4.1 Arhitektura



Slika 4.1: Diagram podatkovnih tokov razločevanja entitet

Proces razločevanja entitet je sestavljen iz treh podprocesov, kot kaže slika 4.1: identificiranje pojavnih oblik, sestavljanje entitet iz pojavnih oblik in nazadnje samo razločevanje teh entitet s pomočjo predznanja. Ta diplomska naloga se osredotoča na tretji podproces in metode, ki so tam uporabljene. Podrobnejša slika arhitekture, ki upošteva tudi vse uporabljene metode, je na voljo v dodatku 1.

4.2 Določanje pomena kot razločevanje entitet

Da lahko na problem določanja pomena pojavne oblike gledamo kot na problem razločevanja entitet, moramo najprej razložiti postopek prevedbe, ki je za to potreben.

Če želimo besedilo obravnavati kot množico entitet, je prvi potrebni korak identifikacija potencialnih entitet v besedilu. Ker pa za razločevanje entitet potrebujemo čim boljše informacije o entiteti iz besedila, za zaznavanje pojavnih oblik uporabimo postopek prepoznavanja imen, ki zaznana imena tudi

klasificira v ustrezne tipe, kot so na primer oseba, organizacija ali lokacija. V ta namen smo uporabili Stanford Named Entity Recognizer [33], ki s pomočjo strojno naučenega klasifikatorja identificira imena v besedilu. Posledično to pomeni, da lahko ena pojavna oblika ustreza več potencialnim entitetam, hkrati pa je lahko potencialna entiteta poimenovana z več pojavnimi oblikami v besedilu.

V splošnem je sicer poljuben n-gram v besedilu tudi pojavna oblika entitete, če v predznanju obstaja entiteta, poimenovana s takšno besedno zvezo. Torej bi lahko z enostavnim pregledovanjem vseh n-gramov in poizvedovanjem po predznanju prišli do entitet, vendar v tem primeru ne bi imeli podatka o klasifikaciji tipa entitete, ki se pojavlja v besedilu.

Preden začnemo z dejanskim iskanjem potencialnih entitet iz predznanja, lahko že na samem besedilu združimo nekatere enake pojavne oblike v skupne entitete, saj pogosto predpostavimo [7], da ima ena pojavna oblika v vsej pojavitvah znotraj posameznega dokumenta vedno enak pomen. V primeru, da pojavno obliko klasificiramo kot ime, spadajo sem tudi okrajšave osebnih imen, kot je npr. "Gospod Novak", ki jo lahko povežemo z "Miha Novak".

Ker pa je pristop s pravili omejen na tovrstne trivialne primere, je potrebno v naslednjem koraku že uporabiti predznanje. Do sedaj lahko dokument že predstavimo kot množico entitet, vendar sedaj še ne vemo, katerim entitetam iz predznanja ustrezajo entitete iz dokumenta.

4.3 Razločevanje entitet

Od tu naprej lahko dokument obravnavamo kot množico entitet, katerim želimo najti ustrezni pomen - entitete iz predznanja.

V osnovni teoriji povezovanja entitet [2] definiramo karakteristike entitet, na podlagi katerih se za vsak par entitet odločamo ali sta enaki ali različni ter ali je sploh možno določiti enakost. Kot karakteristike navadno uporabljajo vsebino atributov, ki jih nato primerjamo z ustreznimi merili podobnosti [34]. Razločevanje entitet je zaporedje odločitev, kjer posamezno entiteto iz dokumenta proglasimo za enako najboljši entiteti iz predznanja glede na kriterij podobnosti na podlagi karakteristik. To ponavljamo, dokler nismo bodisi povezali vseh entitet iz dokumenta z entitetami iz predznanja, bodisi nobena od preostalih entitet iz predznanja ne ustreza kriteriju podobnosti za preostale (nepovezane) entitete iz dokumenta.

Ker je naš model v obliki grafa entitet, lahko karakteristike entitet razporedimo v dva razreda:

- Atributi entitet, kot so imena, opis, številski atributi;
- Relacije do drugih entitet, ki so lahko heterogene;

Na zaporedje odločitev razločevanja lahko gledamo na dva načina:

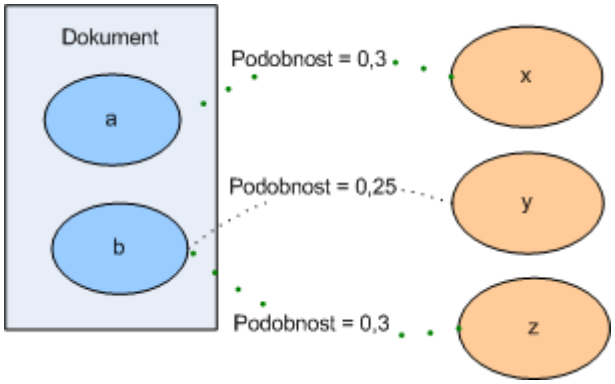
- Posamično razločevanje, kjer je vsaka odločitev neodvisna od ostalih;
- Skupinsko razločevanje, kjer odločitve niso neodvisne in v vsaki odločitvi upoštevamo tudi predhodna razločevanja;

Posebej učinkovito je skupinsko relacijsko razločevanje [9], saj vsak posamezni odločitveni korak spremeni topologijo grafa entitet, kar nam vnese dodatno znanje za nadaljnje korake.

4.4 Posamično razločevanje

4.4.1 Vsebinska podobnost

Ko primerjamo dve entiteti po vsebini, želimo njuno podobnost ovrednotiti sorazmerno s podobnostjo njunih atributov.



Slika 4.2: Posamično razločevanje na podlagi podobnosti - zelene pike označujejo izbrano entiteto

Če ima vsaka entiteta iz dokumenta na voljo več potencialnih kandidatov, kot je razvidno v 4.2, je naša naloga v primeru posamičnega razločevanja oceniti relevantnost vsakega kandidata in na podlagi tega izbrati najboljšega možnega.

Atribute v naši domeni lahko primerjamo z različnimi metrikami:

primerjava enakosti, kjer sta atributa lahko le enaka ali različna;

primerjava nizov po razdalji urejanja, kjer razliko med nizoma ovrednotimo z številom skupnih znakov in številom potrebnih transpozicij, da prvi niz prevedemo v drugega [34];

primerjava daljših nizov v vektorskem prostoru, kjer za mero podobnosti uporabimo kosinusno razdaljo med vektorsko predstavitvijo TF-IDF dimenzij obeh nizov [17].

Naša implementacija obsega prvo in tretjo od teh metrik. Prvo poimenujemo atributna podobnost, ki je definirana z Jaccardovim koeficientom nad množicama atributov entitet.

Naj bodo atributi entitete e_i označeni z atributi_i. Atributi so vse trditve $\langle e_i, p, a\rangle \in KB$, kjer je a element množice dobesednih vrednosti A.

$$\text{atributi}_i = \{\langle p, o\rangle; \langle s, p, o\rangle \in KB \wedge s = e_i \wedge o \in A\} \tag{4.1}$$

Atributna podobnost je torej:

$$\text{podobnost}_{\text{atributi}}(e_i, e_j) = \frac{|\text{atributi}_i \cap \text{atributi}_j|}{|\text{atributi}_i \cup \text{atributi}_j|} \tag{4.2}$$

Besedilno podobnost definiramo kot podobnost med besedili, ki predstavljata entiteti. Naj bo $d_i = \text{Opis}(e_i)$, ki ga predstavimo kot n-dimenzionalni vektor $d_i \in D \subset \Re^n$. V tem vektorju, ki predstavlja vsebino, vsako komponento $n_{i,j}$ vektorja d_i predstavlja frekvenca besede t_j v dokumentu.

$$\text{TF}_{i,j} = \frac{n_{i,j}}{\sum_{t_k \in d_i} n_{k,j}} \tag{4.3}$$

To predstavitev dokumenta imenujemo *vreča besed* (*bag of words*), saj vrstni red besed tu ne igra nobene vloge. [17] tudi ugotavlja, da so uporabnejše tiste besede, ki niso najpogostejše, vendar tudi ne premalo pogoste. To modeliramo z obratno vrednostjo števila dokumentov, v katerih se pojavlja beseda t_j.

$$\text{IDF}_j = \log \frac{|D|}{|d_i \in D; t_j \in d_i|} \tag{4.4}$$

Ko pa dokument želimo prikazati z uteženim vektorjem, je utež komponente t_j v dokumentu d_i je torej določena z $\text{TFIDF}_{i,j}$:

$$\text{TFIDF}_{i,j} = \text{TF}_{i,j} \cdot \text{IDF}_j \tag{4.5}$$

Tako predstavljene dokumente primerjamo s kosinusom prostorskega kota med dvema vektorjema v prostoru, kjer sta d_i in d_j besedilna vektorja entitet e_i in e_j. Naj poudarimo, da je pri posamičnem razločevanju e_i entiteta iz dokumenta, e_j pa iz predznanja. Medtem ko je pri entiteti iz predznanja opis del atributov, pri entiteti iz dokumenta predstavlja besedilni opis kar samo besedilo dokumenta.

$$\text{podobnost}_{\text{besedilna}}(e_i, e_j) = \cos\phi_{i,j} = \frac{d_i \cdot d_j}{\|d_i\| \cdot \|d_j\|} \tag{4.6}$$

Vsebinska podobnost pa je linearna kombinacija atributne podobnosti in besedilne podobnosti:

$$\begin{aligned} \text{podobnost}_{\text{vsebinska}}(e_i, e_j) &= \lambda_{besedilna} \cdot \text{podobnost}_{\text{besedilna}}(e_i, e_j) \\ &+ \lambda_{\text{atributi}} \cdot \text{podobnost}_{\text{atributi}}(e_i, e_j) \end{aligned} \tag{4.7}$$

Algoritem posamičnega ocenjevanja 1 deluje tako, da najprej za vsako entiteto iz dokumenta za vsako pojavno obliko v predznanju najde entitete, ki imajo vrednost rdfs:label enako tej pojavni obliki. S tem definiramo množico entitet, ki so potencialni kandidati. Te kandidate nato ocenimo z vsebinsko podobnostjo, kot smo jo definirali v tem poglavju.

Algorithm 1 Posamično ocenjevanje glede na vsebinsko podobnost

Začetno stanje: Ekstrahirane pojavne oblike P iz besedila $D, P \subset D$.
Začetno stanje: Entitete iz dokumenta $f \in \text{Entitete}_{\text{dokument}}$
Začetno stanje: Predznanje KB, $e \in KB$
  $\text{Entitete}_{\text{kandidati}} \Leftarrow$ prazen seznam
  za vsak $f \in \text{Entitete}_{\text{dokument}}$ **začni**
    za vsak $p \in P, p \in \text{PojavneOblike}_{dokument}(f) \wedge p \in \text{PojavneOblike}_{KB}(e)$ **začni**
      dodaj e v $\text{Entitete}_{\text{kandidati}}$
    konec
  konec

  za vsak $e \in \text{Entitete}_{\text{kandidati}}$ **začni**
    $\text{ocena}_{\text{vsebinska}}(e) = \text{podobnost}_{\text{vsebinska}}(f, e)$
  konec

Razločevanje na tem mestu zaključimo tako, da za vsako entiteto iz dokumenta vzamemo najbolje ocenjenega kandidata iz predznanja, ki je v sliki 4.2

označen z zeleno povezavo. Če želimo nadaljevati s skupinskim razločevanjem, pa kandidate in njihove ocene prenesemo v kasnejše faze.

4.4.2 Relacijska podobnost

Relacijska podobnost izvira iz same topologije grafa entitet. Relacijsko podobnost lahko ovrednotimo z:

- Neposredno povezavo med dvema entitetama;
- Podobnostjo med soseščinama dveh entitet, sorazmerna z velikostjo skupne soseščine;

Ker so v naši domeni povezave heterogene in na tem nivoju še ne predpostavljamo njihovega pomena, je na nek način vseeno potrebno določiti njihovo pomembnost, sicer lahko to informacijo izgubimo. Zato za potrebe razločevanja entitet te povezave utežimo.

Metode za določanje in učenje uteži, primerne za našo domeno, bomo definirali v naslednjih poglavjih.

4.5 Skupinsko razločevanje

Če pri posamičnem razločevanju poskušamo identificirati vsako entiteto posebej, neodvisno od ostalih, pri skupinskem razločevanju ne predpostavimo neodvisnosti med posameznimi odločitvami.

Prednost tega pristopa je, da lahko v kasnejših odločitvah upoštevamo predhodne odločitve o identitetah entitet. Ko se odločimo za pomen neke entitete v besedilu, to pomeni, da ta entiteta iz predznanja nastopa v besedilu. Zaradi posledične spremembe topologije se tudi spremeni relacijska podobnost.

Vzemimo primer, ko imamo dokument, kjer se pojavljajo neznani entiteti, poimenovani z imeni "Elvis"in "Memphis". Prvo je pogosto osebno ime, drugo pa pogosto ime lokacije. V primeru posamičnega razločevanja bi bili prepuščeni hevristikam na podlagi podobnosti posamezne entitete iz dokumenta z kandidati iz predznanja, da bi morda obe izbrali pravilno. Če pa uporabimo skupinsko razločevanje, sprva identificiramo najzanesljivejše kandidate. Če se v tem scenariju vsebinska primerjava odloči, da je dvoumno ime "Elvis"referenca na znanega pevca, v naslednjem koraku skupinskega razločevanja zanesljiveje identificiramo Memphis kot mesto v ameriški zvezni državi Tennessee, kjer je prej identificirani pevec živel, kar je v predznanju v naši domeni izraženo kot utežena povezava med obema entitetama.

Če povzamemo - posamično razločevanje te povezave ne upošteva, zato je zanesljivost nadaljnjih odločitev ponavadi slabša.

Na tem mestu velja poudariti, da zavoljo splošnosti pristopa ne predvidevamo semantike povezave med entitetama, niti ne zahtevamo, da je povezava osnovana na eksplicitni relaciji v predznanju - povezanost entitet lahko izvira tudi iz implicitnih dejstev, kot sta na primer podobnost in sopojavitve.

Skupinsko razločevanje lahko tudi predstavimo s perspektive principa privlačnosti konteksta (Context Attraction Principle), opisan v [8].

Princip pravi, da če imamo pojavno obliko r, ki kaže na entiteto y_j, ki se pojavi v istem kontekstu kot entiteta x, r pa lahko predstavlja katerokoli od entitet $y_1, y_2, \ldots, y_j, \ldots, y_n$, to pomeni, da je x bolj povezan z y_j kot z ostalimi y_i, kjer je $i = 1 \ldots n, i \neq j$.

Če ta princip obrnemo, dobimo sledeče: Če se entiteta x pojavlja v istem kontekstu kot pojavna oblika r, ki predstavlja eno od entitet y_i, in vemo, da je r najtesneje povezan z y_j, potem je od vseh možnih pomenov y_i najverjetnejša izbira $i = j$.

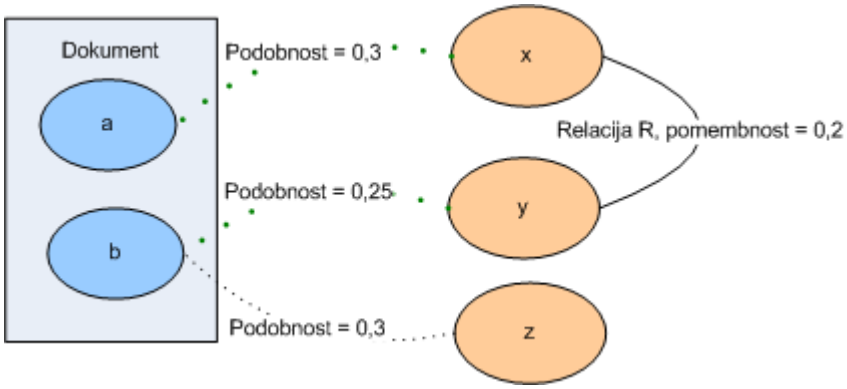
Algorithm 2 Skupinsko relacijsko razločevanje entitet

Začetno stanje: Entitete iz dokumenta $\text{Entitete}_{\text{dokument}}$
Začetno stanje: Entitete iz predznanja $\text{Entitete}_{\text{kandidat}}$
Začetno stanje: Predznanje KB
  Prioritetna vrsta Q
  {Skupinsko ocenjevanje}
  za vsak Potencialni par f, e; $f \in \text{Entitete}_{\text{dokument}} \wedge e \in \text{Entitete}_{\text{kandidat}}$
  začni
    Vstavi $\langle \text{relevantnost}(f, e), f, e \rangle$ v Q
  konec
  Rezultat $\Leftarrow$ prazen seznam
  dokler $Q \neq \emptyset$ **začni**
    $\langle \text{relevantnost}_{f,e}, f, e \rangle \Leftarrow$ vzami iz Q {Predpostavi $f == e$}
    Dodaj $\langle \text{relevantnost}_{f,e}, f, e \rangle \Leftarrow$ v *Rezultat*
    Iz Q odstrani vse kandidate za f
    za vsak $\langle \text{relevantnost}t_{f,e}, f, e \rangle \in Q$ **začni**
      $\text{relevantnost}_{f,e} \Leftarrow \text{relevantnost}(f, e)$
    konec
  konec

Algoritem za skupinsko razločevanje 4.5 deluje tako, da ob vsakem koraku predpostavi enakost para entitete iz dokumenta in entitete iz predznanja z

najvišjo relevantnostjo in glede na to odločitev posodobi vrednost relevantnosti ostalim entitetam v soseščini. Ker predpostavljanje enakosti f in e spremeni topologijo grafa, se spremeni tudi merilo relevantnosti v vseh sosedih e. Zaradi tega jo ob vsaki spremembi stanja posodabljamo. V naslednjih poglavjih si bomo ogledali, kako je moč to relevantnost definirati.

4.5.1 Uporaba relacij



Slika 4.3: Skupinsko razločevanje z eksplicitnimi relacijami

Povezanost med entitetama je v tem primeru izražena eksplicitno v obliki RDF trditve, zapisane v predznanju, kot ponazoruje slika 4.3. V našem primeru predznanje predstavljata združena DBpedia [31] in YAGO [32]. Posamezna RDF trditev je izražena kot trojica osebka, relacije in predmeta. Vzemimo primer, kjer je osebek dbpedia:Elvis Presley v relaciji dbpedia-owl:homeTown s predmetom dbpedia:Memphis,_Tennessee.

Za uporabo v našem modelu razločevanja entitet to relacijo interpretiramo kot povezavo z določeno utežjo. Če so v predznanju relacije homogene in entitete enakih tipov, so posledično tudi vse povezave enako pomembne. V primeru, da ti lastnosti ne veljata, pa je smiselno določati uteži posamezne povezave.

Intuitivno lahko ocenimo, da vse povezave niso enako pomembne. Če bi v našem scenariju v predznanju obstajala le RDF trditev ⟨dbpedia:Elvis Presley, rdf:type, dbpedia-owl:Person⟩, to ne bi bilo pretirano koristno, saj bi najverjetneje vse entitete, imenovane "Elvis"bile na ta način v predznanju opredeljene kot tip dbpedia-owl:Person. Po drugi strani pa je relacija ⟨x, dbpedia-owl:homeTown, dbpedia:Memphis, Tennessee⟩ zelo močan indikator, saj zajema mnogo manjšo množico entitet. Tovrstno lastnost opredelimo kot

selektivnost in njeno vrednost lahko uporabimo kot utež povezave v grafu razločevanja entitet.

Določanje selektivnosti povezave je problem, prisoten tudi v določanju najbolj informativnega podgrafa v danem semantičnem grafu, opisan v [22]. Avtorji so želeli v semantičnem grafu poiskati najmanjši podgraf, ki bi bil še dovolj informativen tako, da izberejo le podmnožico vseh povezav. Za potrebe določanje podmnožice povezav so razvili nekaj metrik, glede na katere to informativnost določajo. Ena od predlaganih metrik, ki je primerna tudi za našo domeno, je selektivnost participacije instanc (*Instance Participation Selectivity*), ki določa, da je selektivnost povezave $\langle s, p, o\rangle$ obratno sorazmerna številu RDF trditev, ki ustrezajo predlogi $\langle type(s), p, type(o)\rangle$, kjer je predikat type($x$) definiran kot relacija rdf:type na entiteti x.

$$\mathrm{IPS}(s, p, o) = \frac{1}{|\pi(\mathrm{type}(s), p, \mathrm{type}(o))|} \tag{4.8}$$

Zaradi bolj uravnoteženih vrednosti IPS smo za namen diplomske naloge definicijo popravili na:

$$\mathrm{IPS}(s, p, o) = \frac{1}{\log(1 + |\pi(\mathrm{type}(s), p, \mathrm{type}(o))|)} \tag{4.9}$$

type(x) predstavlja množico tipov, s katerimi je entiteta x določena.

$$\pi(\mathrm{type}_s, p, \mathrm{type}_o) = \{\langle s, p, o\rangle; \mathrm{type}(s) = \mathrm{type}_s \wedge \mathrm{type}(o) = \mathrm{type}_o\} \tag{4.10}$$

Množica $\pi(\mathrm{type}_s, p, \mathrm{type}_o)$ za tipsko relacijo pa obsega vse relacije v domeni D, kjer je tip osebka enak type_s in tip predmeta enak type_o.

$$\mathrm{type}(x) = \{o \in \langle s, p, o\rangle \,; \langle s, p, o\rangle \in D \wedge s = x \wedge p = \mathrm{rdf{:}type}\} \tag{4.11}$$

Relacijsko povezanost med entitetama e_i in e_j izrazimo z:

$$\mathrm{povezanost}_{\mathrm{neposredna}}(e_i, e_j) = \frac{\sum_{\langle e_i, p, e_j\rangle \in \mathrm{KB}} (\mathrm{IPS}(e_i, p, e_j))}{|\{\langle e_i, p, e_j\rangle \in \mathrm{KB}\}|} \tag{4.12}$$

Ker pa gledamo skupinsko povezanost preko soseščine, pa to izrazimo z:

$$\mathrm{Nbr}(e) = \{f; \mathrm{povezanost}(e, f) \neq 0\} \tag{4.13}$$

$$\mathrm{Nbr}(e_i, e_j) = \mathrm{Nbr}(e_i) \cup \mathrm{Nbr}(e_j) \tag{4.14}$$

$$\begin{aligned} &\mathrm{povezanost}_{\mathrm{sosedska}}(e_i, e_j) = \\ &\frac{\sum_{e_k \in \mathrm{Nbr}_{(e_i,e_j)}} (\mathrm{povezanost}_{(e_i,e_k)} + \mathrm{povezanost}_{(e_j,e_k)})}{|\mathrm{Nbr}_{(e_i,e_j)}|}. \end{aligned} \tag{4.15}$$

To oceno relacijske povezanosti bi lahko posplošili na problem iskanja najkrajše poti skozi graf, kjer je cena poti določena s pogostostjo $|\pi(type(s), p, type(o))|$, kar ustreza $\frac{1}{IPS(s,p,o)}$, vendar se na tem mestu postavlja vprašanje, ali tako dolge povezave dejansko pomenijo smiselno povezanost, zato se raje omejimo na poti dolžine 1 in 2.

V končnih izračunih zato uporabljamo merilo:

$$\begin{aligned} \mathrm{povezanost}_{\mathrm{relacijska}}(e_i, e_j) \quad = \quad & \lambda_1 \cdot \mathrm{povezanost}_{\mathrm{neposredna}}(e_i, e_j) + \\ & \lambda_2 \cdot \mathrm{povezanost}_{\mathrm{sosedska}}(e_i, e_j) \end{aligned} \tag{4.16}$$

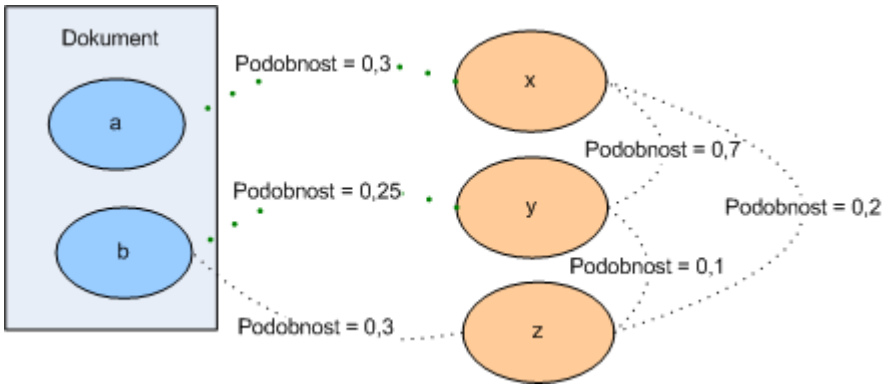
Posledica tega je, da je povezava tipa ⟨dbpedia-owl:Person, dbprop:Origin, dbpedia-owl:Area⟩ manj specifična od ⟨dbpedia-owl:Person, dbprop:Origin, dbpedia-owl:City⟩, kar tudi želimo modelirati.

V tem smo seveda predpostavljali, da imamo za vse entitete v predznanju podatek o tipu, na podlagi katerega lahko določamo selektivnost, potrebujemo pa seveda tudi povezave same - predznanje, ki vsebuje le entitete z atributi za to metodo ni uporabno, vendar za to situacijo obstajajo primernejše metode.

4.5.2 Vsebinska primerjava

V nekaterih situacijah nimamo podanih eksplicitnih relacij med entitetami, vseeno pa jih ne želimo razločevati le posamično. Če so entitete opremljene z opisnimi atributi, s pomočjo katerih je moč izračunati privlačnost med entitetami, lahko definiramo povezanost kar s podobnostjo.

Velja poudariti, da kljub temu da računamo podobnosti tako med entitetami iz dokumenta in tistimi iz predznanja, kot tudi entitetami iz predznanja med seboj, entitet iz predznanja med seboj ne združujemo, saj predznanje obravnavamo kot konsistentno in brez duplikatov. Podobnost je v tem primeru le izraz povezanosti, kot ponazorimo v sliki 4.4.



Slika 4.4: Skupinsko razločevanje z vsebinsko primerjavo

Za vsebinsko primerjavo je najprimernejša predvsem primerjava po opisu z daljšim besedilom, ki poda vsebinski kontekst entitete. Intuicija za tem je, da sta entiteti vsebinsko tesneje povezani, če imata podoben vsebinski kontekst. Ker ne predpostavljamo ničesar o pomenu te povezanosti, razen tega da jo ovrednotimo z utežjo, to ustreza našemu modelu.

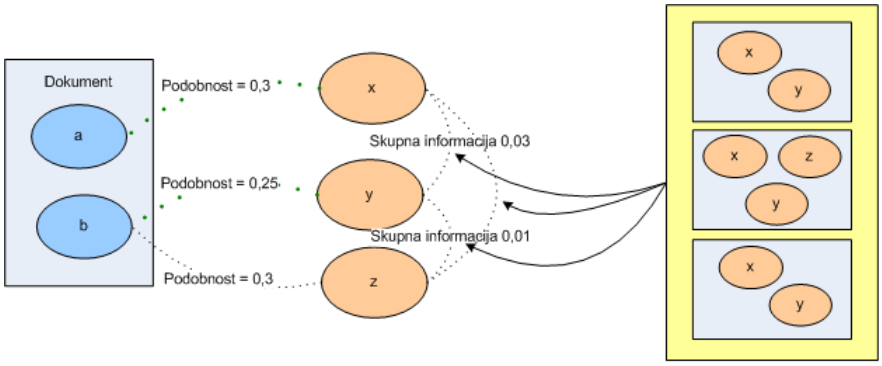
Prednost uporabe vsebinske podobnosti kot povezanosti je, da zadoščajo enaki podatki, kot pri posamični primerjavi. Zaradi tega je primerna tudi za manj strukturirano predznanje, ki ne vsebuje eksplicitnih relacij med entitetami, ima pa vseeno opise entitet. Tudi formula za izračun je enaka formuli za izračun vsebinske podobnosti pri posamičnem razločevanju.

4.5.3 Uporaba sopojavitev

Podatki o sopojavitvah dveh dogodkov se uspešno uporabljajo pri problemih priklica informacij, kot na primer navzkrižni priklic informacij med jeziki [35] in določanje pomena besed [36, 37], ki je problem, soroden razločevanju entitet. Intuicija za uporabo sopojavitev je, da v kolikor se dva dogodka pojavljata skupaj pogosteje kot statistično naključno, bolj je verjetno da sta statistično neodvisna.

Pomembnost sopojvitev modeliramo s količino skupne informacije. Skupna informacija dveh statističnih spremenljivk nam pove, koliko nam znanje o stanju ene spremenljivke zmanjša negotovost stanja druge.

V domeni razločevanja entitet je dogodek definiran kot pojavitev entitete v nekem dokumentu. [38] predlaga uporabo specifične skupne informacije za ocenjevanje povezanosti s sopojavitvami. Specifična skupna informacija (*Specific*



Slika 4.5: Skupinsko razločevanje s statističnim učenjem

Mutual Information[1] - *SI*) sopojavitve entitet x in y v istem dokumentu kvantificira odstopanje med verjetnostjo sopojavitve spremenljivk glede na hkratno pojavitev in sopojavitvijo glede na verjetnost neodvisne pojavitve obeh spremenljivk:

$$SI(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \tag{4.17}$$

V tej enačbi $p(x)$ predstavlja verjetnost pojavitve entitete x, $p(x, y)$ pa verjetnost, da se entiteti x in y pojavita v istem dokumentu.

Vrednost $SI(x, y)$ nato uporabimo kot utež povezanosti entitet x in y pri relacijskem razločevanju.

Če so eksplicitne relacije med entitetami pogosto na voljo za uporabo in enako velja tudi za atributne opise entitet, tega ne moremo vedno trditi za statistiko sopojavitev. Zato v tej diplomski nalogi predlagamo nov postopek za pridobivanje podatkov o sopojavitvah s pomočjo korpusa besedila, kot prikazuje slika 4.5. Če so v korpusu besedila entitete že identificirane, jih je moč trivialno prešteti in s tem pridobiti statistiko sopojavitev parov entitet. Če pa korpus vsebuje zgolj besedilo brez identificiranih entitet, pa jih lahko identificiramo sprva s pomočjo drugih metod razločevanja entitet, ki se ne zanašajo na sopojavitve. To nam da začetno stanje statistike sopojavitev entitet, ki sicer ni nujno popolnoma natančno. Zato se zanašamo na robustnost te metode, da bo te podatke uporabila v naslednji iteraciji razločevanja, kjer lahko doseže višjo natančnost. To implicira, da se z uporabo sopojavitev uporablja v kombinaciji z drugimi metodami razločevanja entitet. Združevanje različnih metod opišemo v naslednjem poglavju.

[1] V literaturi znana tudi kot *Pointwise Mutual Information*

4.6 Združena metoda

V tem poglavju bomo opisali sledeče:

- kako se dopolnjujeta posamično in skupinsko razločevanje;
- kako uporabljati več metod skupinskega razločevanja hkrati.

4.6.1 Združeno posamično in skupinsko razločevanje

Medtem ko je časovna zahtevnost posamičnega razločevanja razreda $O(n)$, kjer je n število entitet iz predznanja, ki so potencialni kandidati za enačenje z entitetami v dokumentu, je skupinsko razločevanje problem razreda $O(m \log m)$, kjer je m število velikost unije soseščin med kandidati iz predznanja. V najslabšem primeru, ko se gre za neusmerjen poln graf, velja $m = \frac{n \cdot (n-1)}{2}$. To pomeni, da je časovna zahtevnost skupinskega razločevanja entitet $O(n^2 \cdot \log n)$.

Zaradi tega je smiselno, da s posamičnim razločevanjem najprej poskušamo zmanjšati množico kandidatov tako, da bo skupinsko razločevanje časovno obvladljivo. V implementaciji to storimo tako, da izločimo vse kandidate, ki so v posamični primerjavi ocenjeni nižje od vnaprej določene meje.

4.6.2 Kombinacija metod skupinskega razločevanja

Vsaka metoda razločevanja entitet kot rezultat vrne seznam izbranih entitet. Če hkrati uporabljamo več možnih metod, se lahko zgodi, da vsaka metoda vrne drugačen seznam. Če želimo kombinirati njihove rezultate, je potrebno za vsako metodo določiti vrednost posameznim kandidatom, nato pa te ocene združiti skupaj in šele na koncu izbrati najboljše kandidate.

Zato pri vseh metodah, tako posamičnih, kot tudi skupinskih, vzamemo za vrednost kandidata povezanost entitete iz predznanja z entiteto iz dokumenta, za končno vrednost pa vzamemo linearno kombinacijo vrednosti posameznih metod.

Če je S množica potencialnih entitet $\{e_i, ...e_j\}$, ki so izbira posamezne metode razločevanja, te kandidate v ovrednotimo z povprečno skupinsko podobnostjo z ostalimi kandidati.

Množico povezav med pari entitet lahko opišemo kot utežene povezave neusmerjenega grafa, kjer uteži povezanost(e_i, e_j) predstavljajo eno izmed treh predlaganih meril povezanosti.

$$\text{povezave} = \{(p, e_i, e_j) \in \Re \times S \times S); p_{ej} = \text{povezanost}(e_i, e_j)\} \tag{4.18}$$

V primeru uporabe vsebinske podobnosti:

$$\text{povezanost}(e_i, e_j) = \text{podobnost}_{\text{vsebinska}}(e_i, e_j) \tag{4.19}$$

V primeru uporabe eksplicitnih relacij:

$$\text{povezanost}(e_i, e_j) = \text{podobnost}_{\text{relacijska}}(e_i, e_j) \tag{4.20}$$

V primeru uporabe sopojavitev:

$$\text{povezanost}(e_i, e_j) = SI(e_i, e_j) \tag{4.21}$$

Ocena posamezne entitete e_i med kandidati S, glede na skupinsko povezanost *povezanost* je povprečna podobnost e_i z drugimi entitetami iz S, s katerimi je povezana:

$$\text{povezanost}(e_i, S) = \frac{\sum_{e_j \in S}(ocena_{\text{povezanost}}(e_i, e_j))}{|S|} \tag{4.22}$$

Posamezno entiteto iz predznanja e, ki jo želimo združiti z entiteto iz dokumenta d nato ovrednotimo z linearno kombinacijo vseh uporabljenih ocen.

$$\begin{aligned} relevantnost_d(e) \quad = \quad & \lambda_1 \cdot \text{podobnost}_{\text{vsebinska}}(e, d) + \\ & \lambda_2 \cdot \text{povezanost}_{\text{vsebinska}}(e, S) + \\ & \lambda_3 \cdot \text{povezanost}_{\text{relacijska}}(e, S) + \\ & \lambda_4 \cdot \text{povezanost}_{SI}(e, S) \end{aligned} \tag{4.23}$$

Poglavje 5

Rezultati

5.1 Metodologija

Za določitev zmogljivosti predlaganih metod smo uporabili meritve natančnosti[1] in priklica[2] glede na izbrano mejo zaupanja v odločitev identificiranja posamezne entitete. Rezultat razločevanja smo primerjali z ročno identificiranimi entitetami, ki so že prisotne kot dodatni metapodatki v uporabljenem korpusu besedila. Ker te ročno identificirane entitete niso podane z enoličnimi označevalniki virov (URI), ki bi jih lahko neposredno preslikali v naše predznanje, to lahko storimo ročno, saj so označene nedvoumno in konsistentno. Ker je množica entitet iz predznanja različna od slovarja ročno označenih entitet, izvajamo vrednotenje na preseku obeh množic.

Končne rezultate podamo v vrednosti F_α, ki je uteženo harmonično povprečje natančnosti in priklica. Natančnost (*precision*) je definirana kot razmerje med pravilnimi izbranimi entitetami in vsemi izbranimi, priklic (*recall*) pa razmerje med pravilnimi izbranimi in vsemi pravilnimi entitetami.

$$F_\alpha = \frac{(1+\alpha) \cdot \text{natančnost} \cdot \text{priklic}}{\alpha \cdot \text{natančnost} + \text{priklic}} \tag{5.1}$$

V praksi se najpogosteje uporablja merilo F_1, torej F_α z $\alpha = 1$, ki natančnost in priklic vrednoti enakovredno. Ker želimo natančnost obravnavati kot pomembneje kot priklic, v meritvah podamo tudi rezultate za $\alpha = 0.2$, ki natančnost uteži petkrat močneje kot priklic.

Iz uporabljenega korpusa New York Times [30], opisanega v poglavju 3., smo skupaj uporabili 39953 člankov, od katerih smo za jih namen meritev

[1] Precision
[2] Recall

ročno označili 79. Teh 79 člankov vsebuje 945 entitet iz predznanja, služi pa testiranju metod.

Ker so metode posamičnega razločevanja, skupinskega z vsebinsko primerjavo in skupinskega z relacijami nenadzorovane, ne potrebujejo učnih podatkov. Kvaliteto rezultato ocenimo s primerjavo z ročno ocenjenimi članki.

Za razliko od ostalih metod, je metoda razločevanja z uporabo sopojavitev nadzorovana. V ta namen ne uporabimo dela testnih primerov, saj jih za statistično učenje potrebujemo bistveno več, kot jih lahko ročno označimo. Zato raje uporabimo samodejno označene dokumente. Entitete v teh dokumentih identificiramo s pomočjo skupinskega razločevanja s semantičnimi relacijami nad preostalim delom korpusa, ki ni uporabljen za testiranje. Ker želimo ohranjati natančnost do največje mere, smo na koncu dokumente označili le s tistimi entitami, ki so bile ocenjene dovolj visoko. Uporabili smo enak prag, ki je v testiranju prinesel 97% natančnost in 45% priklic. Metodo razločevanja entitete s pomočjo relacij smo izbrali, ker ima ta metoda najvišjo natančnost, kot je razvidno v sledečih tabelah.

79 člankov, na katerih smo ročno identificirali entitete smo uporabili za testno množico, medtem ko je ostali del korpusa predstavljal učno množico, na kateri smo šteli sopojavitve entitet.

Implementacijo vseh obravnavanih metod smo tudi preizkusil in dobili sledeče rezultate za posamezne metode. Parametre metod smo določili glede na maksimalno dobljeno vrednost $F_{0.2}$. Za preverjanje pravilnosti rezultatov smo uporabili prej omenjenih 79 dokumentov, glede na katere smo nato izračunali natančnost in priklic.

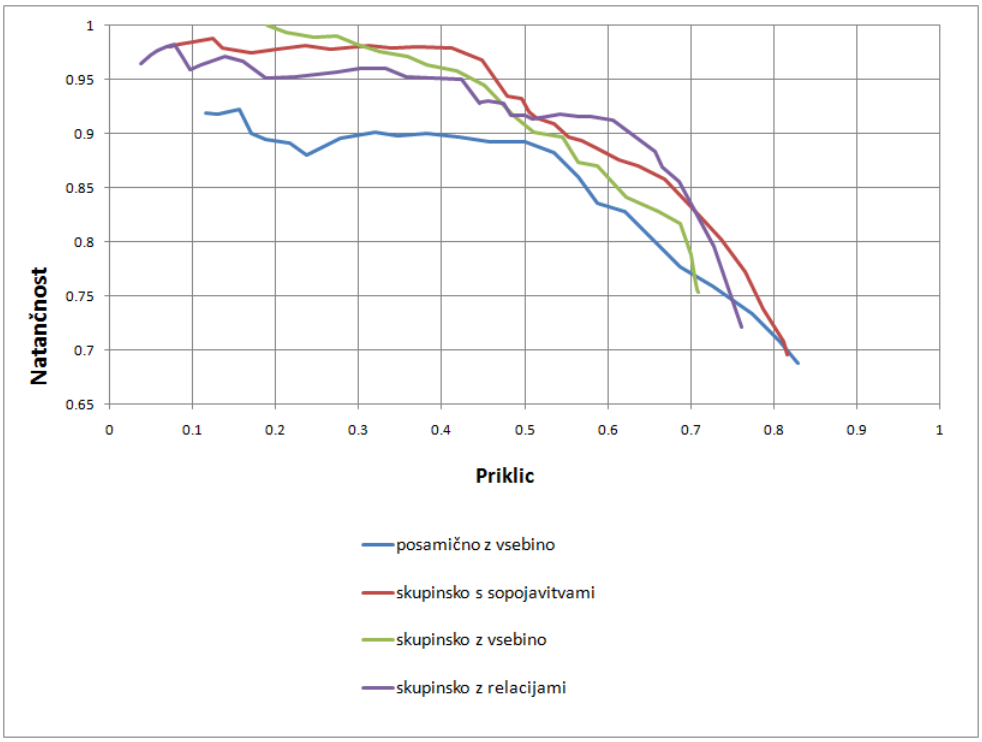
5.2 Rezultati

Metoda za oceno	vir povezanosti	λ utež
Posamično		1.0
Skupinsko	vsebinska	0.9
Skupinsko	sopojavitve	0.05
Skupinsko	relacije	2.0

Tabela 5.1: Rezultati za $F_{1.0}$ in $F_{0.2}$

Uteži rezultatov metod smo določili eksperimentalno, ugotovljene vrednosti

podamo v tabeli 5.1. Vrednosti λ koeficientov so odvisne od predznanja in korpusa besedila. Ker ocene same po sebi niso normalizirane, so odstopanja v velikostnih razredih posledica različnih velikostnih razredov samih ocen. Izrazit primer tega so sopojavitve, kjer specifična skupna informacija ni omejena na interval med 0 in 1, kot je pri povezanosti z vsebinsko podobnostjo. Namen teh uteži je ta, da so posamezne ocene podane v primerljivem velikostnem razredu.



Slika 5.1: Rezultati vseh pristopov

V rezultatih vidimo, da nam dodatne informacije pomagajo doseči boljšo natančnost, vendar ne nujno bolši F_1 kot je videti v 5.2. Skupinske metode razločevanja entitet imajo šibko točko. Ker so posamezne odločitve med seboj soodvisne, ima lahko zgodnja napačna odločitev razločevanja slab vpliv na naslednje odločitve. Posledica tega je, da so v dokumentih, kjer uporabljamo skupinsko razločevanje, entitete bodisi zelo pravilne, bodisi zelo napačne. Zaradi tega je natančnost v manj strogem delovanju lahko celo pri nekaterih dokumentih slabša, kar privede do

Za metode skupinskega razločevanja vidimo, da delujejo odlično pri strožjem razločevanju. V 5.3 vidimo natančnost in priklic pri najvišjem doseženem $F_0.2$, kjer skupinski metodi vsebinske in relacijske povezanosti dosežeta višjo

Metoda	Vir povezanosti	max $F_{1.0}$	max $F_{0.2}$
Posamično		0.754	0.833
Skupinsko	Vsebinska	0.746	0.859
Skupinsko	Sopojavitve	0.769	0.876
Skupinsko	Relacije	0.760	0.841

Tabela 5.2: Rezultati za $F_{1.0}$ in $F_{0.2}$

Metoda	Vir povezanosti	Natančnost max $F_{0.2}$	Priklic max $F_{0.2}$
Posamično		0.839	0.563
Skupinsko	Vsebinska	0.897	0.545
Skupinsko	Sopojavitve	0.857	0.632
Skupinsko	Relacije	0.920	0.556

Tabela 5.3: Rezultati natančnosti in priklica za $F_{0.2}$

natančnost, vendar rahlo slabši priklic, pri sopojavitvah pa vidimo izboljšavo priklica, vendar manj izrazito izboljšavo natančnosti. V tem pogledu posebej izstopa skupinsko razločevanje z relacijami z doseženo 92,0% natančnostjo.

Kot je razvidno v tabeli5.4, pri zahtevani 90% natančnosti skupinske metode odrežejo z bistveno bolje kot posamična, ki doseže 38,1% priklica, medtem ko skupinsko razločevanje z vsebinsko podobnostjo doseže 52,5%, sopojavitve 54,3%, najboljše pa so eksplicitne relacije z 63,7%. Glede na ta rezultat lahko trdimo, da je cilj delovanja pri visoki natančnosti dosegljiv. Po drugi strani pa se natančnost skupinskih metod pri višjem priklicu približuje natančnosti posamičnega razločevanja, kar posledično prinese precej podobne F_1 vrednosti za vse metode.

Med različnimi viri povezanosti, ki smo jih uporabili pri skupinskem razločevanju, kot najboljše izstopa relacijsko predznanje. Izkaže se, da eksplicitne relacije navsezadnje zanesljiveje ocenjujejo povezanost med entitetami kot implicitna povezanost v obliki sopojavitev ali vsebinske podobnosti. Sopojavitve so se izkazale za manj stabilne, ker se zanašajo na dobro natančnost faze razločevanja, ki te sopojavitve identificira. Zaradi tega se napake, storjene v predhodnem razločevanju entitet prenesejo tudi v izračun skupne informacije

Metoda	Vir povezanosti	Priklic pri 90% natančnosti
Posamično		0.381
Skupinsko	Vsebinska	0.525
Skupinsko	Sopojavitve	0.543
Skupinsko	Relacije	0.637

Tabela 5.4: Rezultati priklica pri zahtevani 90% natančnosti

para entitet, kar privede do novih neželenih razločevanj. Skupinska vsebinska primerjava tudi deluje bistveno bolje kot posamična vsebinska primerjava.

Poglavje 6

Zaključek

V sklopu diplomske naloge smo definirali ogrodje za skupinsko razločevanje na podlagi povezanosti. Kot primere povezanosti smo uporabili tri različne metode skupinskega razločevanja, kjer je vsaka primerna za določen tip predznanja, ki ga imamo v dani situaciji na voljo. Med temi metodami je nova metoda razločevanja z določanjem pomembnosti povezav, ki sicer že znane hevristike uporablja v namen razločevanja entitet.

Skupinsko razločevanje z vsebinsko primerjavo je torej možno uporabljati povsod, kjer lahko uporabljamo posamično razločevanje. Prikazan je način izkoriščanja korpusa besedila za skupinsko razločevanja in način, ki uporablja bogatejše relacijsko znanje.

Za izboljšavo bi si lahko pomagali z uporabo strojnega učenja tudi na drugih mestih. Lahko bi ga uporabili za določanje pomembnosti relacij namesto izračuna selektivnosti, uporaba pa je možna tudi za določanje pomembnosti posameznih ocen v zadnjem koraku izračuna skupne ocene.

Ker v nalogi opisana rešitev predstavlja tudi zaključeno funkcionalnost, je na voljo tudi kot spletna storitev [1].

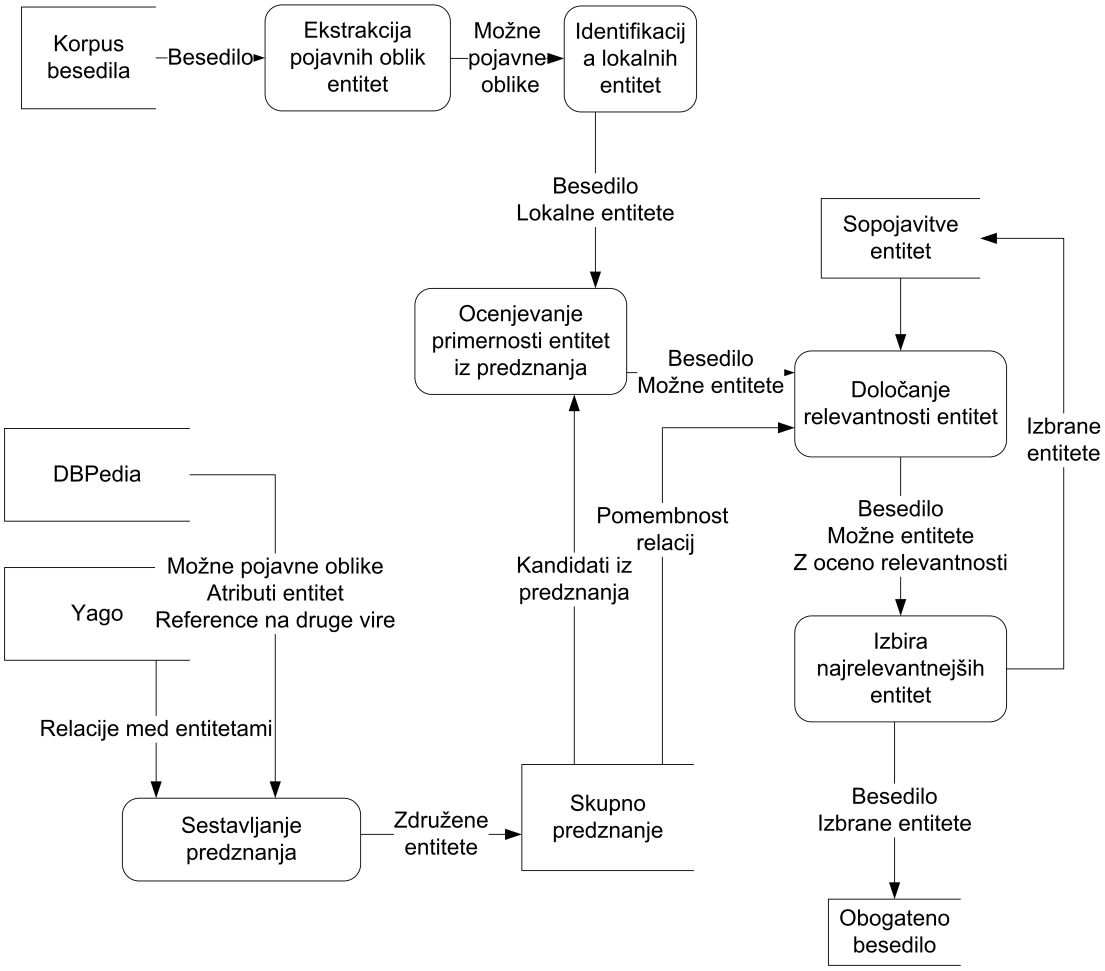
Predlagana rešitev zmožna razločevanja entitet iz besedila je pomemben del ekstrakcije znanja iz besedila. Naslednji nivo ekstrakcije znanja bi bil, če bi poleg entitet iz besedila identificirali tudi relacije, ki se pojavljajo med njimi. Iz teh novonastalih relacij med entitetami lahko izbiramo nove trditve, s katerimi dopolnjujemo predznanje. S tem bi bil vzpostavljen krog, s čimer bi obstoječe znanje lahko izkoristili za iskanje novega znanja. Ta postopek s seboj prinese nove izzive, predvsem na področju izbire ustreznih trditev na podlagi primernosti za vnos v bazo znanja, kot jih obravnava [39].

Uporaba tovrstne tehnologije pa je lahko koristna tudi v druge namene.

[1] http://enrycher.ijs.si

Semantično izražene entitete v omogočajo integracijo in interoperabilnost z zunanjimi podatkovnimi viri [40]. Tudi vizualizacija vsebine besedila v obliki, opisani v [41], je lahko s to dodatno informacijo bogatejša. S tovrstnim semantičnim bogatenjem informacij si lahko olajšamo tudi izzive iskanja informacij, kot na primer sledenje vsebinskim povezavam med dokumenti [42]. Vsem tem primerom uporabe je skupno to, da je njihovo uspešno delovanje odvisno od visoke zmogljivosti razločevanja entitet.

.1 Dodatek A



Slika 1: Arhitektura sistema za razločevanje entitet

Arhitektura prikazuje sistem, katerega sestavljajo 4 različni primeri uporabe:

Sestavljanje predznanja , ki posamezne podatkovne zbirke združi v enotno in jo prevede v obliko, primerno za razločevanje entitet.

Razločevanje entitet , kjer sprva v dokumentih identificiramo entitete, ugotovimo koreference med posameznimi entitetami v dokumentu in jih šele nato poskusimo razločiti s pomočjo predznanja.

Štetje sopojavitev , kjer izbrane entitete posameznega dokumenta uporabimo za statistično učenje, kjer ta podatek izkoristimo pri skupinskem razločevanju s pomočjo sopojavitev.

Izračun pomembnosti relacij je proces, ki ga je za dano predznanje potrebno zagnati le enkrat. Rezultat tega nato lahko uporabimo pri skupinskem razločevanju s pomočjo relacij.

Slike

Tabele

List of Algorithms

Literatura

[1] D. Mladenić, "Text Mining: Machine Learning on Documents," *Encyclopedia of Data Warehousing and Mining*, pp. 1109–1112, 2006.

[2] I. Fellegi and A. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, pp. 1183–1210, 1969.

[3] L. Haas, R. Miller, B. Niswonger, M. Roth, P. Schwarz, and E. Wimmers, "Transforming heterogeneous data with database middleware: Beyond integration," *IEEE Data Engineering Bulletin*, vol. 22, no. 1, pp. 31–36, 1999.

[4] W. Winkler, "The state of record linkage and current research problems," *Statistical Research Division, US Bureau of the Census, Wachington, DC*, 1999.

[5] S. Tejada, C. Knoblock, and S. Minton, "Learning object identification rules for information integration," *Information Systems*, vol. 26, no. 8, pp. 607–633, 2001.

[6] A. Elmagarmid, P. Ipeirotis, and V. Verykios, "Duplicate Record Detection: A Survey," 2006.

[7] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pp. 189–196, Association for Computational Linguistics Morristown, NJ, USA, 1995.

[8] D. Kalashnikov and S. Mehrotra, "A probabilistic model for entity disambiguation using relationships," in *SIAM International Conference on Data Mining (SDM). Newport Beach, California*, pp. 21–23, 2005.

[9] I. Bhattacharya and L. Getoor, "Collective entity resolution in relational data," 2007.

[10] H. Schütze, "Automatic word sense discrimination," *Computational Linguistics*, vol. 24, no. 1, pp. 97–123, 1998.

[11] R. Bunescu and M. Pasca, "Using encyclopedic knowledge for named entity disambiguation," in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3–7, 2006.

[12] S. Cucerzan, "Large-scale named entity disambiguation based on Wikipedia data," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 708–716, 2007.

[13] G. Klyne, J. Carroll, and B. McBride, "Resource description framework (RDF): Concepts and abstract syntax," *W3C recommendation*, vol. 10, 2004.

[14] C. Bizer and A. Seaborne, "D2RQ-treating non-RDF databases as virtual RDF graphs," in *Proceedings of the 3rd International Semantic Web Conference (ISWC2004)*, 2004.

[15] A. McCallum, "Information extraction: Distilling structured data from unstructured text," *Queue*, vol. 3, no. 9, pp. 48–57, 2005.

[16] L. Lloyd, V. Bhagwan, D. Gruhl, and A. Tomkins, "Disambiguation of references to individuals," *IBM Research Report*, 2005.

[17] G. Salton, A. Wong, and C. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[18] R. Mihalcea, "Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 411–418, Association for Computational Linguistics Morristown, NJ, USA, 2005.

[19] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[20] P. Singla and P. Domingos, "Entity resolution with markov logic," in *Proceedings of the Sixth IEEE International Conference on Data Mining*, pp. 572–582, 2006.

[21] Z. Chen, D. Kalashnikov, and S. Mehrotra, "Adaptive graphical approach to entity resolution," in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pp. 204–213, ACM New York, NY, USA, 2007.

[22] C. Ramakrishnan, W. H. Milnor, M. Perry, and A. P. Sheth, "Discovering informative connection subgraphs in multi-relational graphs," *SIGKDD Explor. Newsl.*, vol. 7, no. 2, pp. 56–63, 2005.

[23] M. Grineva, M. Grinev, and D. Lizorkin, "Extracting key terms from noisy and multitheme documents," in *Proceedings of the 18th international conference on World wide web*, pp. 661–670, ACM New York, NY, USA, 2009.

[24] X. Li, P. Morie, and D. Roth, "Semantic integration in text: From ambiguous names to identifiable entities," *AI Magazine. Special Issue on Semantic Integration*, vol. 26, no. 1, pp. 45–58, 2005.

[25] R. Bunescu, R. Mooney, A. Ramani, and E. Marcotte, "Integrating cooccurrence statistics with information extraction for robust retrieval of protein interactions from Medline," in *Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL*, vol. 6, pp. 49–56, 2006.

[26] S. Overell, J. Magalhaes, and S. Ruger, "Place disambiguation with cooccurrence models," in *CLEF 2006 Workshop, Working notes*, 2006.

[27] A. Yates and O. Etzioni, "Unsupervised resolution of objects and relations on the Web," in *Proceedings of NAACL HLT*, pp. 121–130, 2007.

[28] I. Bhattacharya and L. Getoor, "Entity resolution in graphs," *Mining Graph Data*, p. 311, 2006.

[29] D. V. Kalashnikov and S. Mehrotra, "Domain-independent data cleaning via analysis of entity-relationship graph," *ACM Trans. Database Syst.*, vol. 31, no. 2, pp. 716–767, 2006.

[30] E. Sandhaus, "The New York Times Annotated Corpus," 2008.

[31] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," *Lecture Notes in Computer Science*, vol. 4825, p. 722, 2007.

[32] F. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web*, pp. 697–706, ACM New York, NY, USA, 2007.

[33] J. Finkel, T. Grenager, and C. Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling," *Ann Arbor*, vol. 100, 2005.

[34] W. Cohen, P. Ravikumar, and S. Fienberg, "A comparison of string distance metrics for name-matching tasks," in *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*, 2003.

[35] M. Jang, S. Myaeng, and S. Park, "Using mutual information to resolve query translation ambiguities and query term weighting," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 223–229, Association for Computational Linguistics Morristown, NJ, USA, 1999.

[36] K. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[37] H. Li and N. Abe, "Word clustering and disambiguation based on co-occurrence data," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pp. 749–755, Association for Computational Linguistics Morristown, NJ, USA, 1998.

[38] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press, 1999.

[39] F. M. Suchanek, M. Sozio, and G. Weikum, "Sofie: a self-organizing framework for information extraction," in *WWW '09: Proceedings of the 18th international conference on World wide web*, (New York, NY, USA), pp. 631–640, ACM, 2009.

[40] S. Decker, S. Melnik, F. Van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, and I. Horrocks, "The semantic web: The roles of XML and RDF," *IEEE Internet computing*, vol. 4, no. 5, pp. 63–73, 2000.

[41] B. Fortuna, M. Grobelnik, and D. Mladenić, "Visualization of text document corpus," *Special Issue: Hot Topics in European Agent Research I Guest Editors: Andrea Omicini*, vol. 29, pp. 497–502, 2005.

[42] T. Štajner and M. Grobelnik, "Story link detection wih entity resolution," in *Semantic Search*, 2009.