Višejezično izdvajanje citata iz novinskih članaka

Sarajlić, Jelena

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2019./2020.

Jelena Sarajlić

Višejezično izdvajanje citata iz novinskih članaka

Završni rad

Mentor: prof. dr. sc. Nives Mikelić Preradović

Zagreb, srpanj 2020.

Izjava o akademskoj čestitosti

Izjavljujem i svojim potpisom potvrđujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

(potpis)

Sadržaj

1. Uvod

S napretkom tehnologije, svakim se danom količina informacija znatno povećava. Zbog inflacije informacija više nego ikad javljaju se lažne vijesti i dezinformacije te raste potreba za filtracijom vijesti, ustanovljivanjem njihove točnosti i sustavima koji omogućavaju njihovo brzo i lako pretraživanje.

Izdvajanje citata iz novinskih članaka jedan je od pristupa ovom problemu koji omogućava istraživačima uvid u, među ostalim, to kako različiti mediji prikazuju političare (Salway, Meurer, Hofland i Reigem, 2017), analizu ponašanja javnih osoba (Stoyanova, Yalamov i Koeva, 2016) kao i provjeru točnosti i konzistentnosti pojedinih izvora (Newell, Cowlishaw i Man, 2018). Višejezičnost je važna za izdvajanje citata kako bi se omogućilo praćenje vijesti onako kako ih prikazuju različite države, politički sistemi i slično.

Ovaj se rad bavi višejezičnim izdvajanjem citata iz novinskih članaka. Izdvajanje citata postupak je koji se može po potrebi podijeliti na više različitih koraka, a najčešće se sastoji od ekstrakcije sadržaja citata, detekcije govornika i glagola koji uvodi citat te pridruživanja govornika i glagola koji uvodi citat sadržaju citata. Predstavljaju se i uspoređuju različite metode i alati za izvršavanje ovog zadatka s fokusom na više različitih jezika. Razrješavanje koreferencije nije neophodan, ali je važan dio potpunog alata za izdvajanje citata te se stoga u ovom radu nalazi i opis tome namijenjenih metoda. Rad pruža pregled nekih od sustava za izdvajanje citata te predstavlja i probleme na koji svaki od tih sustava može naići. Na koncu se opisuje metoda ručne anotacije citata iz dijela tekstova SETimes korpusa i problemi koji se pritom javljaju.

2. Terminološka razrada

Prema Hrvatskom jezičnom portalu[1], citat je „izvadak iz teksta koji se točno prenosi, navodi od riječi do riječi (…); navod“.

Školski rječnik hrvatskog jezika[2] preusmjerava korisnika s natuknice „citat“ na natuknicu „navod“, čija definicija glasi: „tekst ili dio teksta doslovno prenesen iz koje veće cjeline“.

Mrežno izdanje Hrvatske enciklopedije[3] navodi da je citat: „doslovan navod riječi, rečenice ili ulomka preuzet iz jednoga djela i uvršten u drugo“.

Kada se u svakodnevnom govoru spominju citati, govornici hrvatskog jezika najčešće prvo pomisle na citat omeđen navodnicima koji se prenosi točno onako kako je izrečen. U ovom će se radu pod terminom „citat“ raspravljati i o nekim drugim oblicima navođenja, a navođenje i citiranje smatraju se istoznačnicama. Razlog tome jest činjenica da se u novinskim člancima nečiji iskaz prenosi i na druge načine, a ti se dijelovi teksta i dalje smatraju citatima. Krestel, Bergler i Witte (2008) smatraju da je funkcija izravnih i neizravnih citata jednaka, a razlikuju se po tome je li iskaz identično prenesen ili parafraziran, odnosno sažet. Prema Newell i sur. (2018), u samo 30% navođenja nečijeg iskaza u novinskim člancima koriste se citati omeđeni navodnicima. Pareti, O'Keefe, Konstas, Curran i Koprinska (2013) ustanovili su da se u korpusima koje su koristili za izradu svog alata citati omeđeni navodnicima pojavljuju u 30% i 52% slučajeva navođenja.

Iako većina autora koristi jednaku ili vrlo sličnu terminologiju, neki autori vrste citata razlikuju na sebi svojstven način, poput Weiser i Watrin (2012) koji uvode distinkciju između označenog i neoznačenog presenog govora umjesto izravnih/neizravnih citata.

Osim izravnih i neizravnih citata, neki jezici (poput njemačkog) u novinskim člancima koriste i slobodnu formu citiranja, koja se može opisati kao neizravno citiranje neke osobe uz slobodnu interpretaciju osobe koja ju citira (Krestel i sur., 2008).

---

[1] Dostupno na: http://hjp.znanje.hr/

[2] Dostupno na: http://rjecnik.hr/

[3] Dostupno na: https://enciklopedija.hr/

U ovom će se radu koristiti podjela prema O'Keefe, Pareti, Curran, Koprinska i Honnibal (2012) na izravne, neizravne i mješovite citate.

Izravni su citati strogo omeđeni navodnicima na njihovom početku i kraju te se koriste za doslovno prenošenje iskaza, kao u primjeru (1a).

Neizravni citati ne sadržavaju navodnike, već parafraziraju ono što je rečeno, kao u primjeru (1b).

Mješoviti citati kombinacija su izravnih i neizravnih citata. Djelomično parafraziraju, ali djelomično i doslovno prenose riječi, kao u primjeru (1c).

(1a) Ana je rekla: „Petar je simpatičan.“

(1b) Ana je rekla da joj je Petar simpatičan.

(1c) Ana je rekla da je Petar „simpatičan“.

3. Izdvajanje citata

„Citat se tipično sastoji od tri elementa: izvor, glagol koji uvodi citat i jedan ili više sadržajnih raspona.“ (Newell i sur., 2018, str. 1)

Za potrebe ovog rada potpuno izdvojenim citatom smatra se citat koji je u cijelosti izoliran iz zadanog teksta te su mu pridruženi govornik i glagol kojim je citat uveden u tekst.

Izdvajanje citata može se podijeliti na ekstrakciju, odnosno detekciju dijelova citata (govornika, sadržaja, glagola koji uvodi citat) i na pridruživanje citata. Pridruživanje citata podrazumijeva povezivanje odgovarajućeg govornika i glagola koji uvodi citat sa sadržajem citata. Prema de La Clergerie i sur. (2011) citat je relevantan samo kada je povezan s osobom koja ga je izrekla, a po potrebi se izdvojenom citatu mogu nadodati i ostale značajke poput lokacije.

Ovo poglavlje podijeljeno je na izdvajanje izravnih, neizravnih i mješovitih citata. Pristupi izdvajanja različitim vrstama citata razlikuju se zbog njihove strukture, no postupci detekcije i pridruživanja govornika i glagola koji uvodi citat uglavnom se mogu primijeniti na sve vrste citata. U ovom su se poglavlju pristupi pokušali otprilike podijeliti prema vrsti citata koje izdvajaju, no valja imati na umu kako navedeni pristupi, zapažanja, kao ni rezultati ne moraju biti isključivi samo za jednu vrstu citata.

3.1. Izdvajanje izravnih citata

Većina se ranih radova na temu izdvajanja citata doticala samo ekstrakcije izravnih citata budući da ih je jednostavno detektirati zahvaljujući njihovoj potpunoj omeđenosti navodnicima (Newell i sur., 2018).

Iako se pristupi ovom procesu uvelike razlikuju, radi preglednosti rada podijeliti ćemo ih u tri koraka: detekcija navodnika i označavanje sadržaja, detekcija glagola koji uvodi citat te detekcija autora/govornika citata.

Pristupi i alati opisani u ovom potpoglavlju ne odnose se nužno isključivo na izravne citate, no smatraju se primjenjivima i prikladnima.

3.1.1. Detekcija navodnika/sadržaja citata

Prvi korak u ekstrakciji izravnih citata najčešće je detekcija navodnika i označavanje sadržaja unutar navodnika kao citat. Pri radu s višejezičnim izdvajanjem citata treba imati na umu da različiti jezici koriste različite navodnike, pa tako Poliquen, Steinberger i Best (2007) zapažaju sljedeće razlike u tipografskim oznakama za navođenje:

- **'**…**'**
- **`**…**`**
- **„**…**„** (nizozemski)
- **«**…**»** (francuski)
- **“**…**“**
- **"**…**"**
- **‘**…**‘**
- **---** … (švedska konvencija citiranja nalaže da se na početku citirane rečenice stave jedna ili dvije crtice)

Salway, Meurer, Hofland i Reigem (2017) u svom su radu ustanovili da je u njihovom uzorku tekstova iz norveških novina 630 od 690 izravnih citata uvedeno s dvije crtice na početku citirane rečenice umjesto navodnicima.

Stoyanova i sur. (2016) primijetili su za bugarski jezik uporabu crtice u novom redu prilikom citiranja, no napominju kako je takva pojava rijetka u domeni vijesti.

Prema hrvatskom se pravopisu[4] prilikom navođenja koriste navodnici u obliku **„**…**“** te »...«, no zabilježeno je i korištenje navodnika u obliku **-**…**-**.

Vrste tipografskih oznaka citiranja ne moraju biti isključivo rezervirane ili vezane za neki jezik.

Krestel i sur. (2008) navode da izravni citati mogu imati različite sintaktičke uloge, poput zavisne surečenice[5] ili subjektne dopune[6] te da sadržaj citata može biti razdvojen ostalim rečeničnim dijelovima (primjer 2) ili se on može rastezati na više rečenica.

[4] Dostupno na http://pravopis.hr/
[5] Autori koriste izraz „subordinate clause“. (Krestel i sur., 2008, str. 4)
[6] Autori koriste izraz „subject complement“. (Krestel i sur., 2008, str. 4)

(2) „To je za nas jako važno“ rekao je predsjednik, „zbog činjenice da ima velik utjecaj na ekonomiju“.

De La Clergerie i sur. (2011) vrše ekstrakciju teksta omeđenog navodnicima pomoću simboličnih obrazaca. Takav tekst nazivaju *verbatim*, a uz njega detektiraju i glagole govorenja te imena/nazive u blizini *verbatim*-a. Ovo je prvi od tri koraka u njihovom sustavu, a tek se u zadnjem koraku citat izdvaja uzimajući u obzir rezultate parsiranja i *verbatim* ekstrakcije. Pretpostavka jest da, gledajući sintaktičko stablo rečenice, citati mogu biti objekt glagola koji ih uvodi. Rečenica se označava kao potpuni citat ako je glagolski čvor naveden u unaprijed sastavljenom popisu, a objekt glagola označava se kao sadržaj citata. Njihov alat prepoznaje izravne i mještovite citate.

Liang, Dhillon i Koperski (2010) provjeravaju pristunost navodnika tek u sklopu drugog koraka njihove metode ekstrakcije citata.

3.1.2. Detekcija i pridruživanje glagola koji uvodi citat

Poliquen i sur. (2007) navode kako bez detekcije glagola koji može uvesti citat njihov algoritam uopće neće izdvojiti citat.

Alharabi, Desclés i Suh (2010) imaju šire shvaćanje ovog koraka te u svom radu traže bilo kakav lingvistički marker koji se odnosi na čin govorenja. Ovo proširenje omogućava im detekciju i ostalih vrsta leksema, poput pridjeva i priloga, a koji su važni za razumijevanje okolnosti u kojima je citat izrečen. Njihov alat na koncu kategorizira citate u semantičke kategorije.

Ostali se autori ipak opredjeljuju samo za glagole, koje nazivaju *reporting verbs* (glagoli koji se koriste za ponavljanje nečijeg iskaza) (Poliquen i sur., 2007), *verb-cue* (glagol-signalizator, signalizira uvođenje citata) (Newell i sur., 2018), *quotation verbs* (glagoli citiranja) (Sagot, Danlos i Stern, 2010) ili jednostavno *speech verbs* (glagoli govorenja) (Weiser i Watrin, 2012).

Sagot i sur. (2010) za potrebe automatske ekstrakcije citata oformili su leksikon glagola koji mogu uvoditi citate te ga uključili u *Lefff,* sintaktički leksikon francuskog jezika. Prilikom izrade leksikona ustanovili su da se čak 49% glagola koji uvode citate

može pojavljivati u obrascu citiranja gdje je glagol nakon ili u sredini izravnog citata, ali ne i uz neizravne citate.[7]

Krestel i sur. (2008) koriste alat za označavanje glagola koji uvode citate temeljen na konačnom pretvorniku. Alat pretražuje tekst i uspoređuje korijene glagola s korijenima glagola koji sačinjavaju unaprijed sastavljen popis glagola koji uvode citate. Ukoliko dođe do preklapanja između glagola u rečenici i glagola na popisu, alat označava glagol u rečenici kao glagol koji uvodi citat.

Salway i sur. (2017) sastavili su listu od 64 glagola prikupljenih iz njihova uzorka tekstova temeljenog na člancima pisanim norveškim jezikom. Budući da oni svoj pristup temelje samo na citatima poznatih političara, glagoli koji su stavljani u listu su oni koji se nalaze u kotekstu uz imena političara. Popis je proširen sinonimima glagola. Prema njihovoj analizi, mali broj glagola se najviše koristi prilikom citiranja. Ovi autori u svom pristupu ne rade strogu razliku između vrsta citata.

De La Clergerie i sur. (2011) ručno su sastavili popis od 114 glagola koji mogu uvoditi citate. Njihov pristup oslanja se na sintaktičku analizu rečenice te se glagoli u tekstu uspoređuju s glagolima s popisa. Rečenica se označava kao segment prenesenog govora ukoliko se njezin glagolski čvor nalazi na popisu glagola koji mogu uvoditi citat.

Prvi korak u ekstrakciji citata u radu od Liang i sur. (2010) je uspoređivanje glagola u tekstu s unaprijed određenom listom glagola koji mogu uvesti ili implicirati citat. Citati-kandidati određuju se na temelju prisutnosti prikladnog glagola i postojanja navodnika.

3.1.3. Detekcija i pridruživanje govornika

Prema smjernicama novinske agencije AFP (Agence FrancePresse) iskazi čiji je autor kolektiv ne smatraju se citatima (de La Clergerie, 2011).

Ostali autori uglavnom i kolektive smatraju validnim autorima citata (primjerice, Stoyanova i sur., 2016).

Poliquen i sur. (2007) za izvršavanje zadaće detekcije govornika koriste dvije kategorije: oznake za osobe te osobna imena. Oznake za osobe su najčešće titule, ali

[7] Ovdje je važno naglasiti da Sagot i sur. (2010) razlikuju izravne citate kojima glagol koji ih uvodi prethodi (skraćeno ih zovu DI) i izravne citate u kojima je glagol koji ih uvodi u srednjoj ili finalnoj poziciji (dakle, dolazi u sredini citata ili nakon citata; skraćeno ih zovu DP). Njihova se opaska o mogućnosti pojavljivanju glagola odnosi isključivo na DP.

mogu biti i nacionalnosti, starost i slično. Za potrebe prepoznavanja tih oznaka izgrađen je poseban popis. Osobna imena automatski se prepoznavaju unutar sustava na kojemu autori grade alat (NewsExplorer).

De La Clergerie i sur. (2011) označavaju sintaktički subjekt glagolskog čvora kao autora citata.

Liang i sur. (2010) također uzimaju subjekt glagola kojega su prepoznali kao glagol koji može uvoditi citat za autora citata.

Salway i sur. (2017) definirali su zatvoren skup od 99 norveških najvažnijih i najrecentnijih političara koji su mogli biti autori određenih citata. Nakon toga su iz korpusa uzimali samo one tekstove koji te političare spominju, uz relevantan glagol i navodnike u blizini.

3.1.4. Različiti pristupi ekstrakciji izravnih citata

Poliquen i sur. (2007) navode tri opća pravila (3a-3c) kojima su se služili za detekciju citata, te još tri jezično specifična pravila (3d-3f)[8]:

(3a) *tipografska-oznaka* CITAT *tipografska-oznaka* (,) *glagol* (modifikator) (član) (oznaka osobe) *ime*

(3b) *ime* (, do 60 znakova ,) glagol (: ili riječ koja uz glagol uvodi citat) *tipografska-oznaka* CITAT *tipografska-oznaka*

(3c) *tipografska-oznaka* CITAT *tipografska-oznaka* (; ili ,) (oznaka osobe) ime (modifikator) *glagol*

(3d) *tipografska-oznaka* CITAT1 – (modifikator) *glagol ime* – CITAT 2 *tipografska-oznaka*

(3e) --- CITAT *glagol* (prilog) (oznaka osobe) *ime*

(3f) *glagol* (oznaka osobe) *ime* (modifikator) *tipografska-oznaka* CITAT *tipografska-oznaka*

Obrasci poput (3c) zapaženi su samo u ruskim i talijanskim tekstovima. Švedska konvencija pisanja citata je označena u (3e), dok je (3f) specifičan obrazac za arapski jezik.

[8] Pravila su prenešena iz izvornika uz minimalnu doradu – korištene su oble umjesto uglatih zagrada; eng. *that* u (3b) je zamijenjeno s „riječ koja uz glagol uvodi citat“; umjesto znaka za logičku operaciju „ili“ u (3b) korištena riječ „ili“ i sl.

Njihov se pristup temelji na gore prikazanim leksičkim obrascima i regularnim izrazima, a svoj su sistem automatske analize novinskih vijesti nazvan NewsExplorer učinili javno dostupnim. U trenutku izdavanja njihovog rada NewsExplorer radio je izdvajanje citata iz novinskih članaka na čak 11 jezika[9]. Korištenje obrazaca pri detekciji citata omogućava im lako daljnje širenje na ostale jezike.

De La Clergerie i sur. (2011) svoj sustav ekstrakcije i pridruživanja citata SAPIENS temelje na lingvističkoj obradi (NewsProcess) kojeg dijele na tri glavna dijela: predobrada (u koje ulazi opojavničenje, prepoznavanje imena i/ili naziva (eng. *named entity recognition, skraćeno* NERC) i ekstrakcija citata), duboko parsiranje i na kraju postobrada. Čini se da SAPIENS detektira izravne i mješovite citate, a radi s francuskim jezikom.

Alharabi i sur. (2010) automatsku anotaciju vrše pomoću sustava utemeljenog na pravilima nazvanog EXCOM. EXCOM ne vrši morfosintaktičku analizu ili prepoznavanje imena i/ili naziva (eng. *named entitiy recognition,* tj. NERC), već se koristi površinskim oblicima određenih lingvističkih markera za anotaciju (Alharabi i Desclés, 2009 prema Alharabi i sur., 2010). Njihov rad opisuje sustav koji, nakon anotacije izravnih citata na arapskom, francuskom i korejskom, iste automatski semantički kategorizira.

3.1.5. Problemi ekstrakcije izravnih citata

Iako naizgled jednostavan proces, kod ekstrakcije izravnih citata javljaju se neki problemi na čije rješavanje treba obratiti pozornost.

Alharabi i sur. (2010) u uzorak su nastojali uključiti što je više moguće „teških“ slučajeva citiranja kako bi najbolje mogli testirati svoj sustav, poput lažnih navodnika (navodnika koji ne uvode citat), samocitiranja i „izmišljenih“ citata (primjerice, rečenice u kondicionalu). Primijetili su tri tipa pogrešaka prilikom testiranja sustava: sustav nije prepoznao neke od citata zbog toga što lingvistički markeri koji indiciraju citat još nisu bili uvedeni u bazu podataka, ekstrakcija navodnika koji označavaju nešto što nisu citati kao citate te ekstrakcija navodnika koji označavaju sekunde kao citate kada su u prisutnosti odgovarajućeg lingvističkog markera. Iako se nisu susreli s takvim

[9] Jezici u kojima je NewsExplorer tada vršio izdvajanje citata su arapski, engleski, francuski, nizozemski, njemački, portugalski, rumunjski, ruski, španjolski, švedski i talijanski.

primjerom, kao potencijalan problem navode ugniježđene citate koji bi mogli uzrokovati da se pravi citat ne ekstrahira u potpunosti, već samo do pojave uvodnih navodnika ugniježđenog citata.

Poliquen i sur. (2007) navode pogreške pri ekstrakciji citata koji su razdvojeni na dva dijela (umetanjem nekog drugog rečeničnog dijela), neprepoznavanje citata čiji je autor spomenut samo zamjenicom te neprepoznavanje citata zbog nepoznatog glagola koji ga uvodi (što je riješeno uvođenjem glagola u popis glagola koji uvode citat).

3.2. Izdvajanje neizravnih citata

Krestel i sur. (2008) navode da je većina prenesenih iskaza u obliku neizravnih citata te da se uz njih često dodaju i kontekstne informacije za bolje razumijevanje okolnosti u kojima je citat izrečen.

Iako u svom radu ne čine razliku između izravnih i neizravnih citata, Salway i sur. (2017) navode da su obavezne (glagolske) dopune koje dolaze s veznikom[10] uglavnom neizravan citat.

Stoyanova i sur. (2016) navode kako se neizravan govor obično izražava u obliku zavisne surečenice.

Postupak ekstrakcije neizravnih citata ima neke dodirne točke s postupkom ekstrakcije izravnih citata, poput pronalaska glagola koji uvodi citat i pronalaska govornika, no detekcija prenešenog iskaza uvelike se razlikuje budući da nema tipografskih oznaka koje jasno odvajaju iskaz od ostatka rečenice.

Pristupi i alati opisani u ovom poglavlju ne odnose se nužno isključivo na neizravne citate, no smatraju se primjenjivima i prikladnima.

3.2.1. Detekcija i pridruživanje glagola koji uvodi citat

Weiser i Watrin (2012) pretpostavili su da isti glagoli koji uvode izravne citate uvode i neizravne. Sagradili su korpus od rečenica koje imaju par navodnika i glagol. Glagole koji se pojavljuju u blizini navodnika izolirali su te napominju da ti glagoli mogu izravno

[10] U svom radu autori spominju "…that complements with a complementizer are mostly indirect speech." (Salway i sur., 2017, str. 295) *Complement* se odnosi na obaveznu glagolsku dopunu, koja obično čini citat, a *complementizer* je bilo koja riječ koja uvodi u subjektnu ili objektnu surečenicu. Dakle, značenje u izvorniku šire je od ovdje prevedenog značenja.

tvoriti rječnik glagola koji uvode citate. Dodaju kako bi se iz tog popisa trebali izbrisati glagoli koji nisu dovoljno frekventni da bi bili relevantni, ali i da frekvencija sama po sebi nije dovoljno dobar kriterij već bi trebalo uvesti kriterij povezanosti glagola s citiranjima.

Pareti i sur. (2013) navode kako su glagoli najčešća vrsta riječi koja uvodi citate – čak 96% citata u jednom od korištenih korpusa uvedeno je glagolom. U svojem radu koriste posebnu komponentu unutar sustava koja je zadužena za prepoznavanje glagola koji uvode citat budući da smatraju da nije moguće osloniti se samo na unaprijed sastavljenu listu glagola govorenja. Tu pretpostavku temelje na činjenici da je u jednom od njihovih korpusa 37,5% citata uvedeno glagolom s jedinstvenom pojavnošću.

Newell i sur. (2018) slijedili su pristup od Pareti i sur. (2013) uz manje promjene nekih od značajki.

3.2.2. Detekcija i pridruživanje govornika

Weiser i Watrin (2012) svoj pristup temelje na sintaktičkim obrascima te ne provode anotaciju osobnih imena, stoga kao govornike mogu prepoznati one pojavnosti u tekstu koje se sastoje od osobnih imena i još jedne riječi te svojom pozicijom u rečenici odgovaraju nekom od sintaktičkih obrazaca.

Pareti i sur. (2013) usporedili su 4 metode za detekciju i pridruživanje govornika svim vrstama citata: jednostavan pristup temeljen na pravilima, metoda koja koristi CRF (*Conditional Random Field*), „binarni MaxEnt klasifikator“ (Pareti i sur., 2013, str. 997) i pristup koji koristi „slijedne značajke utemeljene na zlatnom standardu“ (Pareti i sur., 2013, str. 997). Pokazalo se da je pristup temeljen na zlatnom standardu najbolje prepoznavao govornike, no autori napominju kako ovakav pristup nije moguć u praktičnim uporabama. Pristup koji je realno ostvariv i daje najbolje rezultate jest binarni MaxEnt klasifikator. Ovaj pristup svaki od entiteta označava binarno, kao govornika ili ne-govornika, te se onaj entitet s najvećom vjerojatnošću odabire kao govornik.

Newell i sur. (2018) koriste se dvofaznim pristupom za razrješavanje problema u detekciji sadržaja citata i detekciji govornika citata. Autori koriste termin „izvor“ (*source)* umjesto „govornik“ te se u svojem radu referiraju na raspon sadržaja i raspon izvora, umjesto samo sadržaja i izvora. Obje se faze temelje na binarnom klasifikatoru

maksimalne entropije, kojega koriste i Pareti i sur. (2013). Detekcija govornika („izvora“) utvrđuje udaljenost u riječima između izvora i glagola koji uvodi detekciju, nalaze li se oni u istoj rečenici te i jesu li u istoj surečenici (odvojenoj zarezima)[11].

3.2.3. Detekcija sadržaja citata

Pareti i sur. (2013) i Newell i sur. (2018) oboje koriste IOB metodu za određivanje sadržaja citata.

Pareti i sur. (2013) razvili su dva nadzirana pristupa za ekstrakciju citata – jedan baziran na pojavnicama u kojemu se za svaku pojavnicu određuju značajke, te drugi baziran na sastavnicama, u kojemu se određuju značajke sastavnica. Oba pristupa dijele neke zajedničke značajke, poput kategorija leksičkih, rečeničnih, ovisnosnih značajki i značajki koje se odnose na određeno vanjsko znanje. U pristupu baziranom na pojavnicama, svakoj se pojavnici pridodaje I, O ili B oznaka (kratice temeljene na eng. *inside, outside, beginning,* odnosno „unutar, izvan, početak“)[12]. I oznaka označava pojavnicu unutar citata, B oznaka prvu pojavnicu u citatu, dok O oznaka označava pojavnicu koja je izvan citata. Pridodavanje IOB oznaka čine pomoću zajedničkih značajki nadziranih pristupa, kao i određivanja glagola i njegovih podređenih sastavnica, utvrđivanjem nadređenih pojavnica te sintaktičkim značajkama. Pristup baziran na sastavnicama cijele sastavnice određuje kao citat ili ne-citat, a osim zajedničkih značajki koristi se i rasponom, sintaktičkim čvorovima i kontekstom.

Newell i sur. (2018) koriste klasifikatore za glagol koji uvodi citat, klasifikator za sadržaj te klasifikator raspona izvora kako bi svaku pojavnicu mogli označiti IOB oznakama. Navode kako koriste poseban klasifikator sadržaja, za razliku od prijašnjih radova.

Weiser i Watrin (2012) ustvrdili su da se neizravni citati mogu svrstati u 16 sintaktičkih struktura te cijeli proces izdvajanja citata čine pomoću tih struktura. Njihove strukture nisu leksikalizirane, već samo čine moguće obrasce u kojima se neizravni citati mogu iskazati. Takva gramatika pronalazi previše rečenica koje nisu neizravni citat, stoga predlažu leksikalizaciju obrazaca (primarno uvođenjem glagola koji mogu uvesti citate u svoje obrasce).

---

[11] Autori koriste izraz „*parenthetical clause*“. (Newell i sur., 2018, str. 3)
[12] objašnjenje dostupno na Wikipedia članku o IOB označavanju: https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging)

3.2.4. Problemi ekstrakcije neizravnih citata

Neizravni su citati problematični za računalno prepoznavanje i ekstrakciju zbog nedostatka eksplicitne označenosti navodnicima.

U nekim je slučajevima neizravnog citiranja vrlo teško prepoznati granicu između citata i ostatka rečenice, čak i za čovjeka – tako su Weiser i Watrin (2012) ustanovili da za 17,65% citata koji pripadaju jednom od njihovih uzoraka ni čovjek nije bio u stanju ustanoviti gdje završava citat, a počinje ostatak rečenice. Nadalje, autori su primijetili probleme s višeznačnošću nekih glagola koji su krivo označeni kao glagoli koji uvode citat (poput glagola *dodati*, koji je korišten u matematičkom smislu, ili glagola koji su dijelom frazema), te uvrštavanja izraza koji nisu dio citata u citat. Ovi se problemi mogu izbjeći prilagođavanjem sintaktičkih obrazaca.

Osim razaznavanja granice citata, problematično može biti i odrediti što se može smatrati neizravnim citatom, a što ne. Pri prenošenju iskaza u obliku neizravnog citata podrazumijeva se da autor teksta filtrira i parafrazira nečije riječi te se postavlja pitanja do koje mjere filtracija i promjena iskaza mogu ići, a da se nešto i dalje smatra citatom (Salway i sur., 2017).

Newell i sur. (2018) ustanovili su da u nekim slučajevima njihov klasifikator glagola koji uvode citate uopće ne detektira glagol te smatraju da bi razlog mogao biti različitost jezika korištenog za treniranje sustava i onoga koji se doista koristi u novinskim člancima.

3.3. Izdvajanje mješovitih citata

Mješoviti citati sadržavaju i izravne i neizravne dijelove (Newell i sur., 2018).

Razlog njihove razdvojenosti od izravnih citata jest taj da se izdvajanjem samo dijela citata pod navodnicima ne dobiva potpun prikaz iskaza. Razdvajaju se i od neizravnih zbog toga što sadržavaju izravan dio. Omjeri izravnog i neizravnog dijela citata u mješovitim citatima različiti su od citata do citata.

Ekstrakcija mješovitih citata uglavnom se vrši zajedno s ostalim vrstama citata te se ni jedan rad korišten kao izvor za pisanje ovog završnog rada ne bavi isključivo mješovitim citatima.

Neki ih pristupi izdvajaju zajedno s izravnim citatima (de La Clergerie i sur., 2011) (Fernandes, Motta i Milidiú, 2011), a neki s neizravnim citatima (Pareti i sur., 2013).

Stoyanova i sur. (2016) tretiraju izravne i neizravne dijelove mješovitih citata kao zasebne entitete. Budući da njihov alat u kasnijoj fazi grupira sve citate koji imaju zajedničkog autora, ovakvo tretiranje mješovitih citata pokazalo se dostatnim.

4. Razrješavanje koreferencije

Koreferencija se može definirati kao „postupak odabiranja istog referenta u stvarnom svijetu“ (Mitkov, 2002, str. 5).

Pod razrješavanjem koreferencije najčešće se prvenstveno misli na razrješavanje anafore. Osim anafore, razrješavanje koreferencije uključuje i određivanje povezanosti zamjenica s imenicama i određivanje semantičkih podudarnosti između višestrukih spomena nekog entiteta (Almeida, Almeida i Martins, 2014), kao i identifikaciju i prepoznavanje aliasa i skraćenica (Liang i sur., 2010).

Prema Proleksis enciklopediji[13], anafora u književnosti predstavlja stilističku figuru, a u lingvistici odnošenje na prije spomenut referent.

U kontekstu računalne lingvistike i računalnih znanosti, anaforom se smatra „odnos dvaju lingvističkih elemenata u kojemu je interpretacija jednog od elemenata (anafore) na neki način određena interpretacijom drugog elementa (antecedenta).“ (Huang, 2000, str. 1)

Anafora i imenske sintagme koje joj prethode ili ju slijede zajedno čine lanac koreferencije, odnosno skup entiteta koji se odnose na istog referenta. Testom supstitucije možemo ustanoviti entitete koji su u odnosu koreferencije (Mitkov, 2002).

Računalno razrješavanje koreferencije treba sagledati iz više aspekata. Ljudsko razrješavanje koreferencije sastoji se od uvažavanja kontekstnih, gramatičkih, semantičkih i izvanjezičnih indikacija i znanja (Durrett i Klein, 2013).

Razrješavanje koreferencije važno je pri izdvajanju citata jer se autor nekog citata često ne spominje punim imenom (Liang i sur., 2010).

4.1.1. Različiti pristupi razrješavanju koreferencije

Neki sustavi za izdvajanje citata iz novinskih članaka ne upuštaju se u razrješavanje koreferencije osim na razini spajanja različitih varijanti imena spomenutih u tekstu. Primjer takvog pristupa navode Poliquen i sur. (2007) koji razrješavaju samo različite spomene vlastitih imena na način da u tekstu traže riječi koje počinju velikim slovom i također su dijelom imena.

[13] Dostupno na: https://proleksis.lzmk.hr/

Salway i sur. (2017) u svoj su uzorak tekstova uključili samo članke koji spominju autore iz zatvorenog popisa autora te koreferenciju rješavaju pomoću ručno sastavljene tablice koja sadržava alternativne nazive i informacije o spolu. Ograničavaju se samo na ta dva tipa informacija i ne dotiču se razrješavanja vremenski osjetljivih izraza poput „premijer“.

Newell i sur. (2018) sastavili su zaseban alat za razrješavanje koreferencije, koja dolazi kao posljednji korak u njihovom sustavu za izdvajanje citata. Pomoću jednostavnog pristupa temeljenog na pravilima koji pretražuje sve tekstove i traži preklapanje u imenima ili prezimenima, sva imena pronađena u tekstovima stavljaju se u skupine (klastere). Na taj se način razrješavaju aliasi, skraćenice i različiti spomeni istog osobnog imena. Identifikaciju zamjenica koje su označene kao autori rješavaju pomoću pristupa kojega su predložili Durrett i Klein (2014), čiji je alat homogen skup značajki temeljen na podacima. Tim se alatom vrši samo površinska analiza dokumenta i karakteristika te se pokazao boljim u razrješavanju koreferencije od dva dotad najbolja alata, no semantički aspekt razrješavanja koreferencije ostaje izazovom.

Fernandes i sur. (2011) svaku koreferenciju označavaju indeksnim cijelim brojem koji služi kao indikacija povezanosti s ostalim entitetima. Njihov je korpus portugalskog jezika, GloboQuotes, označen zlatnim standardom za, među ostalim, i koreferencije.

Stoyanova i sur. (2016) sva imena i njihove deskriptore spajaju s kanonskim oblicima. Za razrješavanje anafore primjenjuju set jednostavnih pravila. Prvo pronalaze zamjenice trećeg lica jednine u nominativu koji se nalaze neposredno ispred glagola koji uvode citat. Zatim unatrag traže imenicu koja se sa zamjenicom slaže u morfosintaktičkim svojstvima, a anafora se razrješava samo ukoliko je prva imenica koja odgovara zamjenici *named entity*, odnosno ime ili naziv.

Liang i sur. (2010) razrješavanje koreferencije primjenjuju u sklopu lingvističke i semantičke analize, a prije same ekstrakcije i pridruživanja citata.

Almeida i sur. (2014) predstavili su sustav koji zajednički izvršava operacije ekstrakcije citata i razrješavanja koreferencije te ustanovili da je takav pristup bolji nego kada se koriste dva zasebna pristupa posvećena isključivo jednom od tih dvaju zadataka.

5. Pristupi izdvajanju citata

NewsExplorer sustav je koji ekstrakciju citata vrši na temelju sintaktičkih, odnosno leksičkih obrazaca/pravila. Sastoji se od 7 komponenti, a izdvaja samo neke oblike izravnih citata na 11 različitih jezika. Osim govornika, citata i glagola koji uvode citat, izdvajaju i neke modifikatore. Koriste se novinskim člancima koje prikuplja povezana EMM-NewsExplorer aplikacija (Poliquen i sur., 2007).

Krestel i sur. (2008) za izdvajanje citata koriste dva resursa: alat za označavanje glagola koji uvodi citat te alat za pronalaženje citata. Alat za označavanje glagola koji uvodi citat označava glagole koji su dio unaprijed sastavljenog popisa pomoću konačnog pretvornika. Alat za pronalaženje koristi šest obrazaca kako bi prepoznao preneseni govor, odnosno citat, a implementiran je kao regularna gramatika sastavljena od pravila.

Sarmento i Nunes (2009) predstavljaju Verbatim – sustav za izdvajanje citata koji ih slaže prema temi te prikazuje na web-sučelju. Izdvajanje citata vrše pomoću 19 uzoraka i popisom od 35 govornih činova (eng. *speech acts*), odnosno glagola koji uvode citat. Izdvajaju sve vrste citata iz naslova ili tijela novinskih članaka iz mrežnih stranica novina. Odabrali su pristup pomoću uzoraka zbog toga što on ne zahtijeva semantičku analizu te se može relativno lako postići uz pomoć regularnih izraza. Iako izdvajaju sve vrste citata, svjesni su da njihov sustav ne može izdvojiti većinu citata, odnosno pokazalo se da samo 5% vijesti odgovara njihovim uzorcima. Kako bi izbjegli razrješavanje koreferencije i anafore, izdvajaju samo one citate u kojima je govornik eksplicitno naveden.

Alharabi i sur. (2010) predstavljaju sustav koji automatski izdvaja citate omeđene tipografskim znakovima/navodnicima te ih semantički kategorizira. Ne dotiču se pridruživanja citata niti razrješavanja anafore. Budući da se njihov pristup vodi značajkama *enunciative theory* u lingvistici, autori rade razliku između *modus* i *dictum* u logičkom razlikovanju iskaza, kao i *utterer* i *speaker* u polju autora citata. Njihov sustav izdvaja citate na francuskom, arapskom i korejskom jeziku na temelju pravila.

Liang i sur. (2010) predstavljaju sustav za pretraživanje citata prema njihovim semantičkim karakteristikama u puno različitih načina. Autori daju primjer: „What did <speaker> say about <subject>?“ (Liang i sur., 2010, str. 1) kao jedan od mogućih upita za pretraživanje. Citate esktrahiraju s novinskih portala. Prvi korak je lingvistička

i semantička analiza, koja uključuje razdvajanje rečenica, parsiranje, prepoznavanje entiteta, razrješavanje koreferencije itd. Sljedeći je korak detekcija citata – uspoređivanjem glagola u rečenici s unaprijed sastavljenim popisom glagola koji mogu uvoditi citat; detekcijom navodnika. Temeljem ova dva podkoraka ustanovljuju se citati-kandidati. Detektirani citati se postavljaju kao trojka govornik-glagol-citat. Glagol je onaj glagol koji uvodi citat, prepoznat u koraku detekcije, a uz njega se izdvajaju i priložne oznake i ostali važni kontekstni markeri. Subjekt glagola postaje govornikom, a također mu se pridružuju kontekstno važni markeri poput apozicija. Točan opseg citata određuje se pretraživanjem susjednih rečenica. Svakom se potpuno izdvojenom citatu pridružuju i metapodaci. Čini se da njihov sustav u obzir uzima samo izravne citate; potencijalno i izravne dijelove mješovitih citata.

SAPIENS je „platforma za ekstrakciju citata koja se oslanja na dubinski lingvistički lanac procesuiranja“ (de La Clergerie i sur., 2011, str. 5). Arhitektura platforme sastoji se od površinske i dubinske lingvističke razine. Površinska razina detektira tipografske znakove, tj. navodnike, i glagole koji se nalaze u njihovoj blizini u sklopu predobrade, koja uključuje i standardno prepoznavanje naziva i/ili imena te ekstrakciju sadržaja citata. Ekstrakcija citata vrši se pomoću simboličkih uzoraka. Dubinska se razina može podijeliti na duboko parsiranje i postobradu. U postobradu ulazi razrješavanje anafore te u konačnici izdvajanje citata temeljeno na prijašnjim koracima. Od 28 raspona sadržaja citata za koje je zaključeno da ne bi mogli biti u cijelosti izdvojeni pomoću alata koji se ne koristi parsiranjem, SAPIENS je točno identificirao 21 raspon. Ovaj sustav izdvaja izravne i mješovite citate iz novinskog korpusa koji se u konačnici vizualiziraju na web-stranici (de La Clergerie i sur., 2011).

Fernandes i sur. (2011) predstavljaju sustav za izdvajanje citata temeljen na strojnom učenju. Njihov se sustav bavi portugalskim jezikom, no ističu kako je pristup kao takav neovisan o jeziku. Izdvajanje citata dijele na prepoznavanje i pridruživanje citata. Za svaki od podzadataka napravili su *baseline*, odnosno temeljni sistem sastavljen od ručno izrađenih regularnih izraza. Podzadatak prepoznavanja dijele na prepoznavanje početka, kraja i granica citata. Jednom kada sustav prepozna početak citata, kraj i granice citata heuristički se ustanovljuju. Pridruživanje se vrši zajedno s razrješavanjem koreferencije na način da se anafore označe istim indeksnim cijelim brojem, a govornik i citat također se označuju istim indeksnim cijelim brojem. Sustav izdvaja izravne i mješovite citate, a autori su za potrebe izgradnje sustava od vijesti s

portala globo.com sastavili novi korpus nazvan GloboQuotes i označili ga zlatnim standardom.

Weiser i Watrin (2012) čine razliku između označenog i neoznačenog prenesenog govora, a u svojem se radu fokusiraju na neoznačene citate. Ovu distinkciju preferiraju zbog toga što smatraju da označeni citati uglavnom odgovaraju izravnim citatima, no neki neizravni citati također mogu imati jednaku strukturu. Koriste se korpusom novinskih članaka. Pristup je baziran na pravilima, odnosno 16 sintaktičkih struktura koje neoznačen govor može poprimiti. Pravila nisu leksikalizirana, a autori smatraju kako se leksikalizacijom (prvenstveno glagola koji uvode citat) rezultati mogu poboljšati. Osim govornika, sadržaja citata i glagola koji uvode citat, njihov alat izdvaja i rečenične elemente poput vremenskih ili prostornih informacija.

O'Keefe i sur. (2012) ograničavaju se na izdvajanje izravnih citata i izravnih dijelova mješovitih citata. Prvo postavljaju temelje: za ekstrakciju citata koriste se regularnim izrazima koji traže tekst omeđen tipografskim znakovima, tj. navodnicima. Nakon ekstrakcije slijede koraci za detekciju govornika i glagola koji uvode citat. Ovi se koraci također temelje na pravilima. Nakon što su ustvrdili osnovice izdvajanja citata i uspješno izdvojili jednostavnije slučajeve, uvode dva različita pristupa za određivanje govornika (binarni i n-pristup) te tri različita pristupa dekodiranju niza (*greedy, viterbi* i CRF) koje koriste kako bi ustanovili najbolji pristup. U radu se koriste dvama novinskim korpusima i jednim literarnim. Velik dio rada temeljen je na i uspoređuje se s alatom kojeg su uveli Elson i McKeown za izdvajanje citata u domeni književnosti. Njihov je rad „prva evaluacija velikih dimenzija sistema za pridruživanje citata na novinskim člancima“ (O'Keefe i sur., 2012, str. 798).

Pareti i sur. (2013) predstavljaju prvi sustav za izdvajanje svih vrsta citata. U predobradi vrše opojavničavanje dokumenata, označavaju vrste riječi te normaliziraju navodnike. Grade poseban alat za prepoznavanje glagola koji uvode citate. Uz dva *baseline*, odnosno temeljna pristupa, eksprimentiraju i s dva nazdirana pristupa: jedan baziran na pojavnicama, a drugi na sastavnicama. Za pridruživanje govornika citatu testiraju četiri različita pristupa. Za rad koriste dva korpusa iz domene vijesti.

Almeida i sur. (2014) predstavljaju sistem koji pridruživanje citata i razrješavanje koreferenciju tretira kao jedinstven zadatak. Koriste se stablima koreferencije u kojima „spomeni“ i citati predstavljaju čvorove, a logički je program odgovoran za dekodiranje

pridruživanja i koreferencije. Iako kod ovakvog pristupa nije moguće vidjeti tijek odlučivanja programa (i eventualno nešto ispraviti), no autori to ne smatraju velikim problemom. Na koncu zaključuju da ovakav pristup daje bolje rezultate nego odvojeni pristupi pridruživanju citata i razrješavanju koreferencije, a u obzir uzimaju samo izravne i mješovite citate.

Salway i sur. (2017) od resursa koriste popis od 99 najpoznatijih i (subjektivno) relevantnijih norveških političara, te popis od 64 glagola koji mogu uvoditi citat. Samo glagoli govorenja koji traže obaveznu dopunu su mogli postati dijelom tog popisa. Tekstovi koji spominju puno ime jednog od političara parsirali su se, a ukoliko je u tekstu detektiran jedan od flektivnih oblika glagola s popisa, ekstrahirali su se njegov subjekt i njegova obavezna dopuna. Subjekt se označava kao govornik citata, a obavezna dopuna glagola kao citat. Razlog zbog kojega su se autori odlučili raditi samo s tekstovima koji spominju odabrane političare jest taj da si olakšaju razrješavanje koreferencije pomoću ručno složene tablice koja sadržava ostale varijante imena i informacije o spolu. Pomoću te se tablice različiti spomeni istog političara lako mogu razriješiti, kao i slučajevi u kojima je zamjenica u trećem licu označena kao autor. Uzorak koji autori koriste za razvijanje sustava jest skup svakog 220. teksta koji sadržava ime političara, glagol govorenja i navodnike u blizini. Izdvajaju se sve vrste citata na norveškom jeziku.

Newell i sur. (2018) zasebnim alatima identificiraju izvor citata (govornika), glagol koji uvodi citat te sadržaj ili sadržaje citata. Glagoli se detektiraju pomoću klasifikatora, a govornik i sadržaj citata pomoću pristupa temeljenog na pojavnicama kojega su uveli Pareti i sur. (2013) te CRFSuite-a, alata kojega je predstavio Okazaki (2007) (prema Newell i sur., 2018). Pridruživanje citata radi se na dvije razine, a za svaku je sastavljen poseban alat. Razrješavanje koreferencije još je jedan od zasebnih alata koji kombinira prisup temeljen na pravilima s pristupom predstavljenim u Durrett i sur. (2013) (prema Newell i sur., 2018). Izdvajaju sve vrste citata, a na koncu se vrši analiza sentimenta.

Stoyanova i sur. (2016) koriste jezično specifične resurse za bugarski kako bi izdvajali citate iz novinskih članaka: rječnik glagola, popis obrazaca koje citati mogu poprimiti i rječnik koji povezuje imena s apozicijama ili deskriptorima. Rječnik glagola sastavljen je na temelju bugarskog WordNet-a tako da su se uzeli svi sinonimi i hiponimi glagola koji se može prevesti kao „govoriti“. Uvode sistem bodovanja entiteta koji pomaže

pridruživanju citata. Koriste se skupom tekstova izvučenih iz šest velikih novinskih mrežnih stranica (u trenutku pisanja njihovog članka, no čini se da u trenutku pisanja ovog rada koriste 12 novinskih mrežnih stranica[14]). Izdvajaju citate na bugarskom jeziku.

[14] Prema „Sources“ kartici na njihovoj mrežnoj stranici: http://dcl.bas.bg/quotations/lang/en/

6. Izazovi višejezičnosti

Pri izradi višejezičnih računalnih alata, pa tako i kod sustava za višejezično izdvajanje citata, treba imati na umu da se jezici uvelike razlikuju po svome ustroju: bilo to prema fonotaktičkim, morfosintaktičkim ili drugim pravilima. U obzir treba uzeti prednosti i mane alata koji se koriste jezično (ne)ovisnim značajkama pri izdvajanju citata te lakoću dodavanja takvih značajki.

6.1. Izazovi pri višejezičnom izdvajanju citata

Pristupi poput onog od Poliquen i sur. (2007) i Fernandes i sur. (2011) jezično su neovisni.

Prema Poliquen i sur. (2007) prednost njihovog pristupa je to što je lako proširiv na druge jezike i višejezičnost nije prepreka, ali nedostatak je to što bez sintaktičkog parsiranja alat ne prepoznaje neke rečenične dijelove koji su važni za prepoznavanje citata (npr. modifikatore) koji nisu unaprijed zabilježeni. Autori zapažaju da bi se uvođenjem takvog koraka morali koristiti ili čak izrađivati zasebni parseri za pojedine jezike, što bi posljedično dovelo do otežavanja uvođenja novih jezika.

Kasnije je njihov cjelokupni sustav za ekstrakciju informacija proširen na svahili, a time i alat za izdvajanje citata. Svahili se ne razlikuje po izražavanju citata od ostalih jezika koji su već otprije bili zastupljeni u alatu za izdvajanje citata. Za proširenje je samo trebalo izraditi popis glagola koji uvode citat i popis modifikatora. Ostale značajke, poput prepoznavanja vlastitih imena, već su bile uklopljene u druge alate sustava. Jedno je pravilo dorađeno zbog toga što je zabilježeno nezatvaranje prvog dijela sadržaja citata kada je sadržaj citata razdvojen ostalim rečeničnim dijelovima (Steinberger i sur., 2011).

Alharabi i sur. (2010) uočili su da je u korejskom jeziku prepoznavanje i izdvajanje citata puno lakše nego u arapskom i francuskom jer korejski ima specifične lekseme kojima se naznačuju citati. Što se tiče pozicije markera[15] koji mogu uvesti citate, u francuskom oni mogu biti i na početku, i na kraju, i u sredini citata; u arapskom mogu biti na početku ili na kraju; dok su u korejskom uvijek nakon navodnika.

[15] Autori koriste riječ „marker“ kako bi opisali lekseme koji uvode citat, a oni mogu biti glagoli, imenske fraze, priložne fraze itd.

U njemačkom se, uz „klasični“, koristi i slobodni stil citiranja (u kojemu je citat prožet stavom autora novinskog članka) (Krestel i sur., 2008), što može dovesti do problema pri određivanju granice citata.

Bugarski jezik ima slobodan red riječi pa za njega uporaba nefleksibilnih sintaktičkih obrazaca pri izdvajanju citata ne bi bila povoljna (Stoyanova i sur., 2016). Arapski jezik ima relativno slobodan red riječi (Alharabi i sur., 2010). Hrvatski jezik također ima (relativno) slobodan red riječi.

U francuskom je jeziku obavezna inverzija subjekta kada glagol koji uvodi citat dolazi u surečenici u sredini ili nakon sadržaja citata (Sagot i sur., 2010).

6.2. Izazovi pri razrješavanju koreferencije

U francuskom jeziku svaka rečenica obavezno mora imati subjekt. U onim slučajevima koje bi poznavatelji hrvatske gramatike nazvali „besubjektnom rečenicom“ u francuskom se koristi osobna zamjenica trećeg lica jednine *il* kako bi se upotpunila sintaktička obaveza subjekta, a u tom je slučaju *il* bezličan (Achard, 2015). Primjerice, rečenica „Kiši.“ prevodi se kao „**Il** pleut.“ Engleski jezik također koristi pleonastične zamjenice, točnije pokaznu zamjenicu *it,* a ona se načelno ne smatra anaforičnom jer nije dovoljno specifična (Mitkov, 2002). Pri razrješavanju koreferencije u ova dva jezika nužno je prepoznati koje su pojavnosti tih zamjenica anaforične, a koje pleonastične (pogotovo u francuskom). De La Clergerie i sur. (2011) ovaj korak čine pomoću specijaliziranog sustava ILIMP, a Mitkov (2002) navodi da „automatska identifikacija pleonastičnog *it* u engleskom nije trivijalan zadatak“ (str. 10).

U španjolskom jeziku prvo lice množine pri citiranju može imati značenje govorenja u ime institucije (primjerice, glasnogovornik nekog poduzeća može se koristiti prvim licem množine kada govori u ime poduzeća) ili značenje neodređene osobe, tj. općenito značenje (takav oblik može koristiti novinar kada želi govoriti u ime svih građana). Sukladno ovim različitim značenjima razrješavanju koreferencije prvog lica množine mora se različito pristupiti (Recasens i Martí, 2010).

6.3. Transliteracija

Pri višejezičnom izdvajanju citata transliteracija je neizostavan korak ukoliko citate želimo učiniti dostupnima na jedinstvenom pismu ili ih grupirati prema govorniku.

Poliquen i sur. (2005) opisali su proces i probleme u transliteraciji vlastitih imena. Prepoznavanje vlastitih imena vrši se pomoću riječi koje mogu naznačiti osobu, poput apozicija i titula. U jezicima koji sklanjaju vlastita imena koriste se ručno napisana pravila koja sadrže sve moguće sufikse kako bi se imena mogla prepoznati bez obzira na to u kojem su padežu. Kao glavne probleme pri transliteraciji navode činjenicu da odnosi fonem:grafem u različitim jezicima nisu 1:1 (dakle transliteracija se ne može izvršiti pukom zamjenom grafem za grafem), te da se fonemski invetari jezika međusobno razlikuju. U arapskom se jeziku kratki samoglasnici često ne zapisuju pa je dodavanje nenaznačenih samoglasnika ključan korak u transliteraciji na latinicu. Transliteracija s arapskog, ruskog i grčkog pisma provodi se pomoću ručno izrađenih pravila. Autori vlastita imena standardiziraju tj. normaliziraju prema unaprijed zadanim pravilima te ih zatim uspoređuju, a na koncu grupiraju imena i njihove transliteracije prema kriteriju sličnosti.

7. Anotacija citata novinskih članaka na hrvatskom jeziku

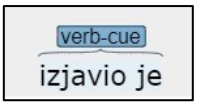
Kao prvi korak sustavu za detekciju i pridruživanje citata, provedena je ručna anotacija citata na hrvatskom jeziku. Tekstovi su preuzeti iz SETimes korpusa, koji se temelji na SETimes novinskoj mrežnoj stranici. U kasnijim bi se fazama anotacija trebala vršiti i na još nekim jezicima, a konačan je cilj izrada alata za određivanje sentimenta iskaza.

7.1. Metoda

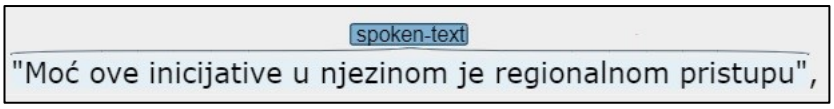
Tekstovi su anotirani pomoću INCEPTION[16] sustava na dvije razine: QuoteFine i QuoteSimple, a fokus ovog poglavlja bit će samo na QuoteFine razini anotacije.

QuoteFine nudi tri moguća tipa oznaka: *speaker*, *verb-cue* i *spoken-text. Speaker* se odnosi na govornika citata (slika 1), *verb-cue* na glagol koji uvodi citat (slika 2), a *spoken-text* na sadržaj citata (slika 3), zajedno s navodnicima koji ga omeđuju. Anotirali su se samo izravni citati i izravni dijelovi mješovitih citata.

Slika 1: Primjer anotacije *speaker*



Slika 2: Primjer anotacije *verb-cue*

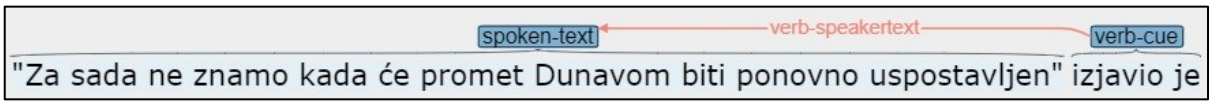


Slika 3: Primjer anotacije *spoken text*

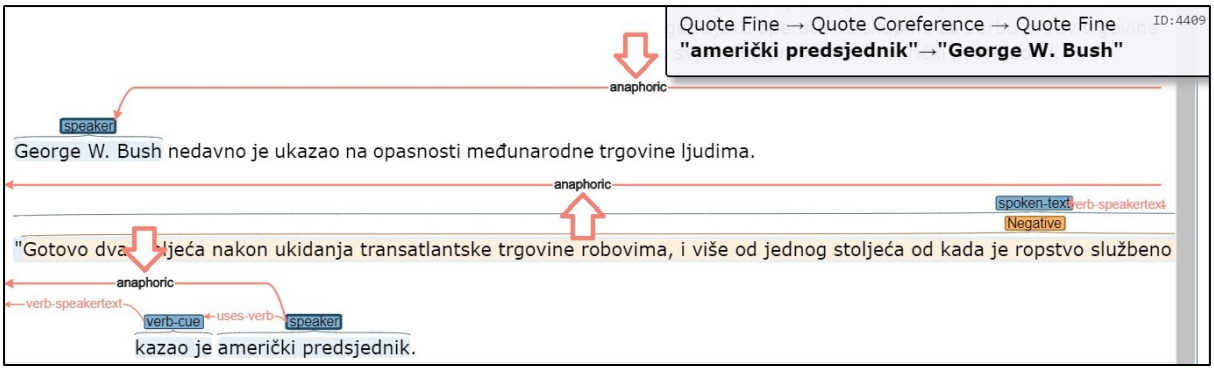
[16] Dostupno na: http://faust.ffzg.hr/inception/

Nakon što se postave ove tri temeljne oznake, one se mogu povezivati pomoću značajke *coreference*. *Coreference* se sastoji od tri vrste oznaka: *uses-verb, verb-speakertext* i *anaphoric. Uses-verb* koristi se za povezivanje govornika i glagola koji uvodi citat (slika 4), *verb-speakertext* za povezivanje glagola koji uvodi citat sa sadržajem citata (slika 5), a *anaphoric* za označavanje entiteta koji su u odnosu anafore (slika 6). *Speaker* i *spoken-text* ne mogu se direktno povezati, već samo indirektno preko *verb-cue*-a, kao što se može vidjeti na slici 7.

Slika 4: Primjer anotacije *uses-verb*



Slika 5: Primjer anotacije *verb-speakertext*



Slika 6: Primjer anotacije *anaphoric* i prikaz veze između entiteta

Slika 7: Prikaz anotiranog citata u cijelosti

Za grubu analizu potencijalnih problema korišteno je prvih 100 tekstova učitanih u sustav INCEPTION koji u sebi sadržavaju navodne znakove. Od 100 dokumenata s navodnim znakovima, 94 dokumenata ima citate, dok 6 dokumenata ima navodnike koji označavaju nešto što nije citat (nazivi poduzeća, dokumenata, prenesena značenja i sl.)

Jedan dokument u prosjeku sadrži 23 rečenice – najkraći dokument ima 5, a najduži 48 rečenica. U 96 dokumenata s citatima prosječno je 13 anotacija, odnosno šest anotacija sadržaja citata, četiri anotacije glagola koji uvode citat i četiri anotacije govornika. Ukupno je anotirano 500 sadržaja citata, 363 glagola koji uvode citat i 354 govornika[17]. Prosječno vrijeme potrebno za anotaciju jednog dokumenta je sedam minuta.

Najčešći glagoli koji uvode citat na ovom uzorku su *kazati, izjaviti* i *reći.* Iz analize su izbačena tri označena *verb-cue*-a jer se radi o glagolskim prilozima, a ne glagolima. Od 360 anotacija glagola koji uvode citat, 110 puta se pojavljuje *kazati*, 84 puta *izjaviti* i 68 puta *reći*. Čak 12 glagola pojavljuje se samo jednom u cijelom uzorku. Treba napomenuti da su se svršeni i nesvršeni oblik nekih glagola analizirali zajedno (istaknuti-isticati, upozoriti-upozoravati). Ukupno je zabilježeno 27 glagola koji uvode citate, a u tablici 1 može se vidjeti postotak pojavljivanja pet najčešćih glagola.

Tablica 1: Prikaz postotka pojavljivanja 5 najčešćih glagola koji uvode citat

GLAGOL	**POSTOTAK POJAVLJIVANJA**
kazati	30,56%
izjaviti	23,33%
reći	18,89%
dodati	6,39%
istaknuti / isticati	4,17%

[17] Brojala se svaka zasebna anotacija, dakle među 354 govornika vjerojatno se više puta pojavljuje isti govornik.

7.2. Problemi

Uz anotaciju, za svaki su dokument bilježene pojavnosti koje bi mogle biti od važnosti pri izgradnji sustava za automatsko izdvajanje citata, ili bi mogle biti pri tom postupku problematične. Takve pojavnosti su različite: od onih koje su čisto tehničke prirode i vezane uz INCEPTION sustav preko konkretnih problema za izdvajanje citata sve do pogrešaka u dokumentima, poput zatipaka.

Budući da nisu sve bilježene pojavnosti od jednake važnosti za predstavljanje u ovom radu, razraditi će se samo oni koji se smatraju važnima. Tablica 2 prikazuje pojavnosti koje će se razraditi te broj dokumenata, odnosno novinskih članaka, u kojima se one pojavljuju.

Tablica 2: Prikaz pojavnosti i broj dokumenata u kojima se pojavljuju

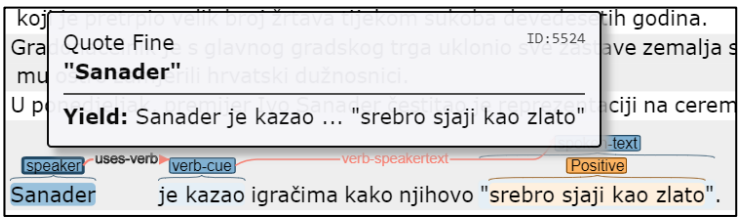
pojavnost	**broj dokumenata u kojima se pojavljuje**
apozicija se nalazi uz govornika ili njegov alternativni spomen	73
govornik se više puta različito spominje kroz tekst	59
indirektno povezan govornik	55
mješoviti citati	52
govornik i/ili glagol koji uvodi citat nalaze se u sredini sadržaja citata	40
navodnici koji označavaju nešto što nije citat	33
citati čiji su autori kolektiv ili su ulomci nekih dokumenata	23
uporaba pasiva	19
govornik nije naveden imenom	15
indirektno povezan glagol koji uvodi citat	13
navođenje bez jasnog govornika i glagola koji uvodi citat	11
govornik nije osoba (već kolektiv i sl.)	9
cross-branching	4
izravni citat koji nije stavljen u navodnike	1

„Idealan“ citat bio bi onaj u kojemu se sadržaj citata nalazi odmah pored govornika i glagola koji uvodi citat; odnosno slučaj u kojemu su govornik, citat i glagol koji uvodi citat u susljednom nizu. Ovaj „strogi“ pristup odabran je kako bi se što više potencijalno problematičnih slučajeva moglo zajedno obuhvatiti pod jednom oznakom pojavnosti te kako ne bi bilo upitno koji slučaj pripada u ovu kategoriju, a koji ne.

Zbog olakšavanja budućeg razrješavanja aliasa i koreferencije, obilježavani su i dokumenti u kojima se apozicija nalazi uz govornika ili njegov alternativni spomen. Važno je istaknuti da se u ovom koraku apozicije koje nisu u ulozi govornika i aliasi nisu označavali anaforičnim odnosom. Najčešće bi se govornik prvi puta u tekstu uveo punim imenom i prezimenom te kasnije samo prezimenom ili, rjeđe, apozicijom. Iz istog razloga obilježavani su dokumenti u kojima se govornik više puta različito spominje kroz tekst. U slučaju različitog spominjanja istog govornika kroz tekst, kao govornik se anotirao onaj spomen bliži sadržaju citata.

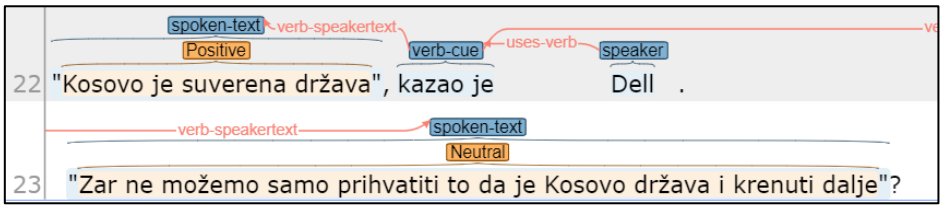
Zbog strogog određenja „idealnog citata“ oznaka indirektno povezan govornik dodjeljivala se svakom govorniku koji nije odmah pored citata ili glagola koji uvodi citat, pa čak i kada je odvojen samo apozicijom.

Mješoviti citati anotirani su ovisno o kontekstu. Sadržaj citata uvijek se anotirao, govornik u slučajevima kada je to bilo moguće (često je bio indirektno povezan s citatom), a glagol koji uvodi citat samo u slučajevima kada se radilo o glagolu govorenja ili komunikacije. Slika 8 prikazuje slučaj mješovitog citata koji je razumljiv i samo uz svoju izravnu komponentu, odnosno prikazuje usporedbu stvarnog teksta i anotacije. Mješoviti citati trenutačno nisu fokus ovog projekta te se stoga njihova analiza nije temeljitije provodila.



Slika 8: Prikaz mješovitog citata čiji bi izravni dio mogao stajati samostalno bez da se uvelike naruši razumljivost

Primijećeno je da se često u sredini prenesenog iskaza uvode govornik i glagol koji uvodi citat, kao što je prikazano na slici 9. Stoga je uvedena oznaka govornik i/ili glagol koji uvodi citat nalaze se u sredini sadržaja citata. Ponekad je samo govornik ili samo glagol koji uvodi citat u sredini sadržaja citata, no ta distinkcija nije smatrana dovoljno relevantnom za bilježenje te su takvi slučajevi zajednički bilježeni pod ovu kategoriju, uz napomenu koji dio se nalazi u sredini sadržaja citata.



Slika 9: Prikaz dijela teksta u kojemu su govornik i glagol koji uvodi citat u sredini sadržaja citata

Navodnici koji označavaju nešto što nije citat obično su emfatični navodnici (od eng. *emphatic quotes*, navodnici koji se stavljaju na nešto što se želi naglasiti ili istaknuti), imena dokumenata i slično. U nekim je dokumentima bilo teško procijeniti jesu li navodnici emfatični ili prenošenje nečijeg iskaza te su se takvi slučajevi označavali ovisno o kontekstu. Slika 10 pruža prikaz takvog slučaja, gdje je teško odrediti je li Bill Gates novootvoreni kompleks nazvao inovacijskim središtem ili su ga tako prozvali novinari i građani.

Bill Gates boravio je u Ateni kako bi obilježio otvaranje novog "inovacijskog središta" tog softverskog diva.

Slika 10: Prikaz slučaja u kojem je teško odrediti jesu li navodnici preneseni govor ili emfatični navodnici

Citati čiji su autor kolektiv ili su ulomci nekih dokumenata obilježavani su kao potencijalan problem zbog smjernice francuske novinske agencije AFP da se citati kolektiva ne smatraju citatima (de La Clergerie i sur., 2011), te zbog toga što se navođenje ulomaka nekih dokumenata isto možda ne bi smatralo konvencionalnim citatima (budući da nemaju „govornika“).

Oznaka pasiv uvedena je zbog nemogućnosti da se u takvim slučajevima anotira govornik. Od 19 dokumenata u kojima je pasiv zabilježen, u čak 16 je zabilježen i citat čiji je autor kolektiv ili u je ulomak nekog dokumenta (84%).

Govornik nije naveden imenom također je jedna od oznaka koje su uvedene radi lakšeg budućeg razrješavanja koreferencije. Tu se često radilo o apozicijama ili zamjenicama, ali zabilježeni su i dokumenti u kojima je citiran anonimni govornik (primjerice, građanin nekog grada i sl.)

S oznakom indirektno povezan glagol koji uvodi citat postupalo se na sličan način kao s oznakom indirektno povezan govornik.

Oznakom navođenje bez jasnog govornika i glagola koji uvodi citat obilježavali su se slučajevi kada je sadržaj citata naveden negdje u sredini teksta, bez eksplicitno i/ili jasno naznačenog govornika i glagola koji uvodi citat. Pretpostavka je da bi čovjek bez većih problema mogao prepoznati tko je autor citata, no računalo bi imalo poteškoća.

Govornik nije osoba oznaka je koja je uvedena iz istog razloga kao i oznaka citati čiji su autor kolektiv ili su ulomci nekih dokumenata.

Cross-branching je pojavnost koja se odnosi na križanje grana u sintaktičkom stablu, što se prema lingvističkoj teoriji neposrednih sastavnica ne bi smjelo događati. No razlog zbog kojega je ova oznaka uvedena nije problematika sastavljanja sintaktičkih stabala, već nemogućnost anotacije rastavljenog glagolskog oblika kao jedinstvenog predikata odnosno *verb-cue*-a. lako je potekla kao čisto tehnička problematika proizašla iz ograničenja sustava za anotaciju, dakako da bi mogla biti problematična i za automatsko izdvajanje citata. Moguće bi rješenje ovakvih slučajeva bilo automatsko dodavanje pomoćnog glagola punoznačnom obliku glagola u prošlom vremenu, budući da se u hrvatskom jeziku iz samog oblika glagola lako može iščitati lice i rod govornika.

Oznaka izravni citat koji nije stavljen u navodnike uvedena je zbog pretpostavke da će postojati slučajevi gdje se izravni citati nisu stavljali u navodnike. Zabilježen je samo jedan dokument u kojemu je to bio slučaj te se pretpostavlja da je riječ o zatipku budući da navodnici nisu stavljeni samo na početku citata, ali jesu na kraju.

U dva su teksta zabilježeni ugniježđeni navodnici.

Problematično je i citiranje „drugog reda“, odnosno slučajevi u kojima se navodi citat koji je izvorno negdje drugdje citiran uz napomenu izvora.

U nastavku anotacije trebalo bi detaljnije razraditi svaki od ovih potencijalnih problema, kao i osmisliti rješenja za ista.

8. Zaključak

Izdvajanje citata iz novinskih članaka postupak je koji može biti samodostatan, ali i korišten u svrhu specifičnijih sustava analize i grupiranja vijesti, ponašanja javnih osoba, javnog mijenja i sl. Riječ je o postupku koji se po potrebi može podijeliti na više različitih koraka, a najčešće se sastoji od ekstrakcije sadržaja citata, detekcije govornika i glagola koji uvodi citat te pridruživanja govornika i glagola koji uvodi citat sadržaju citata.

Pristupi izdvajanju citata iz novinskih članaka uvelike se razlikuju prema tome koju vrstu citata izdvajaju i načinu na koji to čine. Vrste citata najčešće se dijele na izravne citate, koji su omeđeni navodnicima i doslovno prenose nečije riječi, neizravne citate, koji su parafraza izrečenog i uklopljeni su u rečenicu, i mješovite citate, koji imaju neke značajke izravnih i neke značajke neizravnih citata. Mnogi se pristupi oslanjaju na jezično specifične resurse, a oni koji to ne čine riskiraju potencijalno veliku nepreciznost i netočnost sustava. Jezično specifični resursi mogu uključivati rječnike glagola koji uvode citate, sintaktičke parsere i slično. Jezično nespecifični resursi mogu se oslanjati na općenita pravila i obrasce kojima se citati izražavaju na više različitih jezika. Izravni citati puno su jednostavniji za detekciju i ekstrakciju zahvaljujući svojoj omeđenosti navodnicima, dok se za izdvajanje neizravnih citata treba osloniti na detaljniju lingvističku analizu tekstova, bilo to obrada tekstova sintaktičkim alatima ili istraživanje o načinima na koje se neizravni citati mogu pojavljivati. Neki od citiranih autora pokazali su kako se većina citata pojavljuje upravo u neizravnom obliku. Mješoviti se citati najčešće pridodaju ili izravnim ili neizravnim citatima. Način njihova izdvajanja ovisan je o tome kojoj se vrsti citata pridodaju te na koji način autori uopće takve citate žele tretirati. Probleme pri izdvajanju citata mogu činiti, primjerice, višeznačni glagoli, neobične konstrukcije, navodnici koji označavaju preneseno značenje ili nešto drugo što nije citat, ili pak sami resursi korišteni za izdvajanje citata, ukoliko nisu dovoljno detaljno razrađeni. Višejezično izdvajanje citata također sa sobom donosi brojne prepreke, budući da različiti jezici na (vrlo) različit način označavaju i izražavaju citate. Alat koji želi postići dobre rezultate u višejezičnom izdvajanju citata mora imati razvijene jezično specifične resurse za svaki od uključenih jezika ili jezično nespecifične resurse koji bi trebali biti dovoljno ograničeni da ne izazivaju (previše) grešaka pri izdvajanju, a opet dovoljno neograničeni da mogu obuhvatiti citate na više različitih jezika. To je vrlo zahtjevan zadatak, a o tome koliko

je izdvajanje citata, naizgled jednostavan proces, zapravo vrlo kompleksan za računalnu obradu govori podatak da su neki autori zabilježili iznenađujuće velik broj citata kojima ni čovjek nije mogao odrediti točne granice.

Na primjeru ručne anotacije novinskih članaka hrvatskog jezika prikazano je koliko se zapravo implicitnih problema nalazi pri tretiranju citata i navedeno je desetak takvih problema, kao i njihova čestotnost. Neki su problemi specifičniji za hrvatski jezik, poput *cross-branching*-a, dok su ostali problemi pojavnosti koje se mogu pojaviti u bilo kojem jeziku. Provedena je i štura analiza glagola koji uvode citat i njihove čestotnosti.

Budući će se sustavi za izdvajanje citata zasigurno morati koncentrirati na pronalaženje idealne točke između korištenja potpuno jezično specifičnih alata i resursa i potpuno generaliziranih alata i resursa, pogotovo u višejezičnom kontekstu.

9. Literatura

Achard, M. (2015). *Impersonals and other agent defocusing constructions in French* (Vol. 50). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Alharabi, M. i Desclés, J.P. (2009). Opérations de prise en charge énonciative: assertion, médiatif et modalités dans le discours rapporté direct, en arabe et en français. U K. Bogacki, J. Cholewa i A. Rozumko (ur.). *Methods of lexical analysis, theoretical assumptions and practical applications, 17* (ur., 26 str.) Bialystok: Wydawnictwo Uniwersytetu w Bialymstoku. dostupno na: http://lalic.paris-sorbonne.fr/PUBLICATIONS/2009/pologne.pdf

Alrahabi, M., Desclés, J.P. i Suh, J. (2010). Direct Reported Speech in Multilingual Texts: Automatic Annotation and Semantic Categorization. *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference, FLAIRS-23,* 162-167. dostupno na: https://pdfs.semanticscholar.org/8ec6/19aa7810f739f930f4b7a0a965552297447d.pdf

Almeida, M. S. C., Almeida, M. B. i Martins, A. F. T. (2014). A Joint Model for Quotation Attribution and Coreference Resolution. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 39-48. DOI: 10.3115/v1/e14-1005

de La Clergerie, É., Sagot, B., Stern, R., Denis, P., Recourcé, G. i Mignot, V. (2011). Extracting and Visualizing Quotations from News Wires. *Human Language Technology. Challenges for Computer Science and Linguistics Lecture Notes in Computer Science*, 522–532. DOI: 10.1007/978-3-642-20095-3_48

Durrett, G. i Klein, D. (2013). Easy Victories and Uphill Battles in Coreference Resolution. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1971-1982. dostupno na: https://www.aclweb.org/anthology/D13-1203.pdf

Fernandes, W.P.D., Motta, E. i Milidiú, R.L. (2011). Quotation Extraction for Portuguese. *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, 204-208. dostupno na: https://www.aclweb.org/anthology/W11-4527.pdf

Hrvatska enciklopedija, mrežno izdanje. dostupno na: https://enciklopedija.hr/ (datum pristupa 18. lipanj 2020.)

Hrvatski jezični portal. dostupno na: http://hjp.znanje.hr/ (datum pristupa 1. svibanj 2020.)

Hrvatski pravopis. dostupno na: http://pravopis.hr/ (datum pristupa 1. svibanj 2020.)

Huang, Y. (2000). *Anaphora: a cross-linguistic study*. Oxford: Oxford University Press.

Inside–outside–beginning (tagging). dostupno na: https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging) (datum pristupa 18. lipanj 2020.)

Krestel, R., Bergler, S. i Witte, R. (2008). Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 6 str. dostupno na: http://www.lrec-conf.org/proceedings/lrec2008/pdf/718_paper.pdf

Liang, J., Dhillon, N. i Koperski, K. (2010). A large-scale system for annotating and querying quotations in news feeds. *Proceedings of the 3rd International Semantic Search Workshop (SEMSEARCH '10)*, 1–5. DOI: https://doi.org/10.1145/1863879.1863886

Mitkov, R. (2002). *Anaphora Resolution.* Harlow i London, Ujedinjeno Kraljevstvo: Pearson Education.

Newell, C., Cowlishaw, T. i Man, D. (2018). Quote Extraction and Analysis for News. dostupno na: https://research.signal ai.com/assets/RnD_at_the_BBC__and_quotes.pdf

Okazaki, N. (2007). CRFsuite: a fast implementation of conditional random fields (CRFs). dostupno na: http://www.chokkan.org/software/crfsuite/

O'Keefe, T., Pareti, S., Curran, J.R.; Koprinska, I. i Honnibal, M. (2012). A Sequence Labelling Approach to Quote Attribution. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 790–799. dostupno na: https://www.aclweb.org/anthology/D12-1072.pdf

Pareti, S., O'Keefe, T., Konstas, I., Curran, J.R. i Koprinska, I. (2013). Automatically Detecting and Attributing Indirect Quotations. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 989–999. dostupno na: https://www.aclweb.org/anthology/D13-1101.pdf

Pouliquen, B., Steinberger, R., Ignat, C., Temnikova, I. i Widiger, A. (2005). Multilingual person name recognition and transliteration. *Corela. Cognition, représentation, langage*, (HS-2). DOI: 10.4000/corela.1219

Poliquen, B., Steinberger, R. i Best, C. (2007). Automatic Detection of Quotations in Multilingual News. *International Conference 'Recent Advances in Natural Language Processing' – Proceedings*, 487-492. dostupno na: https://pdfs.semanticscholar.org/3979/b0125f77499de215fb29e1c5d8feae5cf476.pdf

Proleksis enciklopedija online. dostupno na: https://proleksis.lzmk.hr/ (datum pristupa 1. svibanj 2020.)

Quotations Retrieval System from Bulgarian Media Content. dostupno na: http://dcl.bas.bg/quotations/lang/en/ (datum pristupa 18. lipanj 2020.)

Recasens, M. i Martí, M. A. (2010). AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language resources and evaluation*, 44(4), 315-345. DOI: 10.1007/s10579-009-9108-x

Sagot, B., Danlos, L. i Stern, R. (2010). A lexicon of french quotation verbs for automatic quotation extraction*. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10),* 294-299. dostupno na: http://www.lrec-conf.org/proceedings/lrec2010/pdf/387_Paper.pdf

Salway, A., Meurer, P., Hofland, K. i Reigem, Ø. (2017). Quote Extraction and Attribution from Norwegian Newspapers. *Proceedings of the 21st Nordic Conference of Computational Linguistics*, 293–297. dostupno na: https://www.aclweb.org/anthology/W17-0241.pdf

Steinberger, R., Ombuya, S., Kabadjov, M., Pouliquen, B., Della Rocca, L., Belyaeva, J., de Paola, M., Ignat, C. i Van der Goot, E. (2011). Expanding a multilingual media monitoring and information extraction tool to a new

language: Swahili*. Language resources and evaluation*, 45(3), 311-330. DOI: https://doi.org/10.1007/s10579-011-9155-y

Stoyanova I., M. Yalamov. i S. Koeva. (2016). Quotation Retrieval System for Bulgarian Media Content. *Proceedings of the Second International Conference Computational Linguistics in Bulgaria 2016*, 64–73. dostupno na: https://www.academia.edu/30840191/Quotation_Retrieval_System_for_Bulgarian_Media_Content

Školski rječnik hrvatskoga jezika. dostupno na: http://rjecnik.hr/ (datum pristupa 1. svibanj 2020.)

Weiser, S. i Watrin, P. (2012). Extraction of unmarked quotations in Newspapers - A study based on direct speech extraction systems*. LREC - International Conference on Language Resources and Evaluation*, 559-562. dostupno na: http://hdl.handle.net/2078.1/120331

Prilozi

glagol	broj pojavljivanja
kazati	110
izjaviti	84
reći	68
dodati	23
istaknuti/isticati	15
navoditi	14
ukazati	6
pojasniti	5
pozvati	*5*
upozoravati/upozoriti	5
zaključiti	5
citirati	2
objasniti	2
opisati	2
prenijeti	*2*
izvijestiti	1
naglasiti	1
najaviti	1
napomenuti	1
odgovoriti	1
primijetiti	1
priopćiti	1
prisjećati se	1
spomenuti	1
tvrditi	1
zanijekati	1
zapitati se	1

Prilog 1: Popis svih glagola koji su označeni kao glagoli koji uvode citat (*verb-cue*). Ukupno je zabilježeno 360 glagola koji su označeni kao *verb-cue*, odnosno glagol koji uvodi citat. Iz analize su izbačena tri *verb-cue*-a jer je riječ o glagolskim prilozima (dakle ukupno su zabilježena 363 *verb-cue*-a), a to su *ukazujući, kazavši, ističući*.

Višejezično izdvajanje citata iz novinskih članaka

Sažetak

Završni rad "Višejezično izdvajanje citata iz novinskih članaka" predstavlja različite pristupe izdvajanju citata na više jezika. Rad opisuje proces izdvajanja citata iz novinskih članaka pisanih na više jezika, kao i postojeće probleme pri tom procesu (detekcija i ekstrakcija sadržaja citata, pridruživanje govornika i glagola koji uvodi citat) kroz više različitih pristupa. Također se prikazuju mogući načini razrješavanja koreferencije, koje su česta pojavnost u novinskim tekstovima. Daje se opis sustava i alata koji izdvajaju sve vrste citata iz desetak različitih jezika. Rad nastoji prikazati konkretne probleme koji pri tom procesu nastaju te usporediti različita moguća rješenja tih problema. Na koncu se opisuje prvi korak u gradnji sustava za izdvajanje citata, a to je ručna anotacija podataka. Anotacija obuhvaća označavanje sadržaja citata i njihovog opsega te označavanje govornika (ili, u slučaju zamjenice ili aliasa, naznačavanje originalnog govornika), kao i glagola koji uvodi citat. Uz opis postupka anotacije, opisuju se i različiti problemi na koje se tijekom anotacije naišlo.

Ključne riječi: izdvajanje citata, višejezičnost, računalna obrada jezika, razrješavanje koreferencije

Multilingual Extraction of Quotes from News Articles

Summary

Bachelor's Thesis “Multilingual Extraction of Quotes from News Articles” presents different approaches to extraction of quotations from news articles in different languages. This thesis describes the process of quotation extraction from news articles written in multiple languages with the existing problems which arise during that process (detecting and extracting the content of the quote, speaker and verb-cue attribution). It also demonstrates different approaches to coreference resolution, which is a frequent occurrence in texts which are in the news domain. A description of systems and tools used to extract all types of quotes from cca. 10 languages is given. This works aims to give an overview of specific problems which can be arise during this process and compare some of the potential solutions. In the end, the first step in building a system which would extract quotes, annotating quotations by hand, is described. The annotation encompasses tagging the content of the quote and its span, tagging the speaker (or the original speaker, in the occurrence of aliases and pronouns), as well as tagging the verb-cue. Along with the process description, a description of the problems which were encountered during the annotation is given.

Key words: quotation extraction, multilinguality, computational processing of language, coreference resolution