Laboratorij za analizu teksta i inženjerstvo znanja – TakeLab

Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Zavod za elektroniku, mikrolektroniku, računalne i inteligentne sustave
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 458

Ekstrakcija događaja i vremenskih relacija u tekstovima na hrvatskome jeziku

Mladen Marović

Zagreb, lipanj 2012.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA DIPLOMSKI RAD PROFILA

Zagreb, 5. ožujka 2012.

DIPLOMSKI ZADATAK br. 458

Pristupnik: **Mladen Marović**
Studij: Računarstvo
Profil: Računarska znanost

Zadatak: **Ekstrakcija događaja i vremenskih relacija u tekstovima na hrvatskome jeziku**

Opis zadatka:

Danas su dostupne goleme količine pisanoga teksta koje predstavljaju velik izvor znanja. Automatska ekstrakcija informacija iz tekstnih podataka, poput ekstrakcije događaja i vremenskih relacija među događajima, omogućava iskorištavanje tog znanja u različitim područjima ljudske djelatnosti. Ekstrakcija događaja i vremenskih relacija netrivijalni su zadatci obrade prirodnog jezika i predmetom su intenzivnog istraživanja.
U okviru diplomskog rada potrebno je proučiti postupke za ekstrakciju događaja i vremenskih relacija temeljene na metodama strojnog učenja. Razraditi postupak za ekstrakciju događaja i vremenskih relacija u tekstovima na hrvatskom jeziku. Provesti označavanje odgovarajućeg tekstnog uzorka i odabrati najprikladnije značajke uzevši u obzir ograničenost jezičnotehnoloških alata za hrvatski jezik. Provesti eksperimentalno vrednovanje točnosti ekstrakcije uporabom različitih metoda strojnog učenja, analizu značajki te detaljnu analizu pogrešaka. Razviti programsku implementaciju postupka korištenjem jedne odabrane metode. Radu priložiti izvorni programski kod, programsku dokumentaciju i ispitne uzorke.

Zadatak uručen pristupniku: 9. ožujka 2012.
Rok za predaju rada: 21. lipnja 2012.

Mentor:

Doc.dr. sc. Jan Šnajder

Djelovođa:

Prof.dr.sc. Domagoj Jakobović

Predsjednik odbora za diplomski rad profila:

Prof.dr.sc. Siniša Srbljić

Zahvaljujem mag. ing. comp. Goranu Glavašu na izradi aplikacija za označavanje događaja i vremenskih relacija te na mnogim savjetima koji su mi pomogli u izradi ovog rada. Također zahvaljujem Damiru Cvetovcu, Latici Čačković, Marijani Marović i Kristijanu Pavloviću na uloženom trudu i ukazanom strpljenju u označavanju i izradi korpusa korištenog u radu.

SADRŽAJ

1. Uvod

Većina ljudskog znanja danas je pohranjena u tekstovnom obliku. Zbog sve bržeg razvoja tehnologije pisani su izvori lako dostupni svima te predstavljaju veliki potencijal za iskorištavanje u različite svrhe. Za automatsku uporabu i učinkovitu obradu pisanih tekstova potreban je razvoj naprednih sustava za pretraživanje informacija i složenih metoda za iskorištavanje pohranjenog znanja. Ovi su problemi u domeni područja pretraživanja informacija (engl. *information retrieval*) i dubinske analize teksta (engl. *text mining*). Pisani tekst sadrži informacije zapisane prirodnim jezikom, stoga se u ovakvim zadacima koriste i mnoga saznanja iz područja obrade prirodnog jezika (engl. *natural language processing*).

Korist koju pisani tekstovi i njihovo iskorištavanje mogu pružiti je višestruka, a primjene su brojne. Osim same pohrane i pretraživanja, česta je uporaba pisanog teksta u raznim zadacima za koje je potrebno razumijevanje teksta slično ljudskom. Tako je često potrebno prepoznati različite entitete u tekstu te njihove uloge i odnose. Znanje o entitetima sadržanim u tekstu računalu pruža dublje razumijevanje semantike teksta od onoga koje se može postići uporabom samo riječi. Ono pak otvara mnoge mogućnosti za rješavanje složenih problema u različitim područjima ljudske djelatnosti, primjerice u područjima koja se bave automatskim odgovaranjem na pitanja (engl. *automatic question answering*), strojnim prevođenjem (engl. *machine translation*) ili sažimanjem dokumenata (engl. *document summarization*).

Otkrivanje semantičkih komponenti poput entiteta, njihovih uloga i odnosa tek je prvi korak u naprednom iskorištavanju bilo kakvih pisanih izvora. Dodatni izazov leži u povezivanju tih komponenti u složenije strukture koje bi mogle predstavljati robusne baze znanja otkrivenog u tekstu. Ovakve baze znanja iskoristive su u daljnjem razumijevanju novih tekstova i izvora, ali i u prezentiranju i generiranju novog znanja.

Cilj ovog rada jest ekstrakcija događaja i vremenskih relacija između događaja u pisanim tekstovima na hrvatskom jeziku. Za ekstrakciju će se koristiti metode strojnog učenja. Razumijevanje teksta na razini događaja i vremenskih relacija riješilo bi neke poteškoće prisutne kod sustava za odgovaranje na pitanja, a koje su prven-

stveno vezane uz nemogućnost povezivanja spomenutih entiteta. Također, dobro izgrađen model temeljen na događajima i vremenskim relacijama otkrivenim u nekim tekstovima mogao bi poslužiti pri zadacima sažimanja istih.

Struktura rada dana je kako slijedi. U idućem poglavlju predstavljen je pregled srodnih radova značajnih za istraživanje provedeno u ovom radu. Opisani su pristupi ekstrakciji događaja i vremenskih relacija te su spomenute spoznaje i saznanja iz tih istraživanja. Zatim slijedi opis teorijskih postavki na kojima se temelji ovaj rad. Dana je definicija događaja i njihovih semantičkih razreda te vremenskih relacija i njihovih vrsta. Četvrto poglavlje bavi se tehničkom stranom istraživanja te daje opis implementacije alata i sustava korištenih pri provođenju eksperimenata. Nakon toga dan je opis postupka izrade korpusa, opis eksperimenata i prikaz dobivenih rezultata. Konačno, rad je zaključen osvrtom na stečene spoznaje i prijedlozima za daljnji rad.

2. Srodni radovi

U ovom je poglavlju dan pregled srodnih istraživanja, njihovih pristupa, metoda, rezultata i spoznaja. Istraživanja vezana uz ekstrakciju događaja obrađuju se u prvom dijelu poglavlja, dok se drugi dio poglavlja bavi radovima na području ekstrakcije vremenskih relacija.

2.1. Ekstrakcija događaja

Početna razmatranja pojave događaja u rečenici pojavila su se u lingvističkim istraživanjima. U istraživanjima poput Vendler (1957); Verkuyl (2005) proučavani su glagoli kao predikati rečenice i događaji koje oni kao nositelji radnje pritom opisuju. Također, razmatrala se i struktura opisanih događaja. Strukturu čine svi oni manji događaji koji su sastavni dio većeg, promatranog događaja. Ako je struktura nekog događaja jednostavna, tada se on ne može podijeliti na manje događaje od kojih je sastavljen. S druge strane, složene strukture mogu se podijeliti na veći broj manjih događaja koji čine cjelinu. Prema novim spoznajama definirana su obilježja predikata na temelju kojih je moguće odrediti smještaj opisanog događaja u vremenu, tj. vrijeme u kojem se on događao. Prema tim obilježjima predikate je moguće podijeliti u sljedeće skupine (Bethard, 2007):

– statički/dinamički (engl. *static/dynamic*) – Statički predikati opisuju događaje koji se ne mijenjaju za vrijeme svog trajanja i koji imaju jednostavnu strukturu. Primjerice, predikat *znati* je statički jer netko nešto zna u bilo kojem trenutku za vrijeme trajanja tog događaja. Također, "znanje" se ne može podijeliti na više manjih aktivnosti koje čine veću. S druge strane, dinamički predikati opisuju događaje koji uključuju neku promjenu i imaju složeniju strukturu. Primjerice, predikat *penjati se* je dinamički predikat sa složenom strukturom koja obuhvaća mnoge jednostavnije događaje, poput *podići ruku*, *uhvatiti se za granu*, *pomaknuti nogu* itd.;

– trajni/trenutni (engl. *durative/punctual*) – Trajni predikati opisuju događaje koji traju neko vrijeme, tj. imaju svoj početak, neko vrijeme tijekom kojeg se odvijaju i završetak. Primjerice, predikat *hodati* je trajni predikat jer on može trajati dulje vrijeme. S druge strane, trenutni predikati opisuju događaje koji se doživljavaju kao trenutni. Njihov početak, trajanje i završetak se ne percipiraju odvojeno u vremenu, već sve prolazi u istom trenutku. Primjeri takvih događaja su *pronaći* i *trepnuti*;

– svršeni/nesvršeni (engl. *telic/atelic*) – Svršeni predikati opisuju događaje koji svojim značenjem obuhvaćaju i završetak događaja, često ostvaren kroz ispunjenje nekog cilja. Primjerice, predikat *pronaći* je svršeni jer opisani događaj obuhvaća potragu za nekim predmetom i konačni pronalazak tog predmeta. S druge strane, nesvršeni predikati opisuju događaje koji ne obuhvaćaju završetak događaja, poput predikata *tražiti*. U hrvatskom jeziku ovo je svojstvo kod glagola izraženo glagolskim vidom.

Opisana obilježja predikata omogućuju klasifikaciju predikata u različite razrede. Primjerice, već u (Vendler, 1957) predložena je sljedeća podjela predikata:

– stanja (engl. *states*) – statički, trajni predikati poput *vjerovati* i *voljeti*;

– aktivnosti (engl. *activities/processes*) – dinamički, trajni, ateličkipredikati poput *hodati* i *pisati seminar*;

– postignuća (engl. *accomplishments*) – dinamički, trajni, svršeni predikati poput *nacrtati krug* i *ozdraviti*;

– ostvarenja (engl. *achievements*) – dinamički, trenutni, ateličkipredikati poput *prepoznati* i *pronaći*.

Druga istraživanja, poput (Pustejovsky, 1991) uvela su proširenja poput hijerarhijske strukture događaja. Unatoč tome, navedena Vendlerova podjela u četiri razreda imala je najveći utjecaj na kasnija lingvistička istraživanja i radove na temu obrade prirodnog jezika.

Prva istraživanja u ovom području koja koriste statističke metode i metode strojnog učena imala su nešto manji opseg. Primjerice, u (Siegel i McKeown, 2000) korištene su metode strojnog učenja za određivanje aspektualnih obilježja glagola. Za potrebe istraživanja korišteno je korišteno je 14 značajki, poput vremena glagola, prisutnosti negacije, prisutnosti subjekta, priloga itd. Za klasifikaciju su korištene logistička regresija, stabla odluke i genetsko programiranje. Istraživanje je provedeno u dva dijela. U prvom dijelu istraživanja proučavana je klasifikacija glagola u stanja i događaje.

Istraživanje je provedeno na 3.224 medicinskih otpusnica, što je obuhvaćalo preko $1.100.000$ riječi s ukupno 1.478 glagola, od čega je njih 739 (634 događaja) bilo iskorišteno za učenje modela, a ostalih 739 (619 događaja) za testiranje. Postignuta je točnost od $93{,}9\%$. U drugom dijelu proučavana je klasifikacija u svršene i nesvršene događaje. Korišteni korpus sastojao se od deset romana, što je obuhvaćalo nešto manje od 850.000 riječi. U te riječi bilo je ubrojeno 615 događaja, od čega je njih 307 (196 svršenih) korišteno za učenje , a 308 (195 svršenih) za testiranje. Postignuta je točnost od $74{,}0\%$. Rezultati istraživanja pokazali su da je korištenjem jednostavnih lingvističkih značajki moguće primijeniti metode strojnog učenja za klasifikaciju događaja u različite razrede, primjerice one dane u (Vendler, 1957).

U svrhu boljeg istraživanja ekstrakcije događaja i vremenskih relacija u tekstu pokrenut je projekt *TimeML*, opisan u (Pustejovsky et al., 2003a). *TimeML* je bogati specifikacijski jezik za označavanje događaja i vremenskih izraza u tekstu. On definira događaje kao trajne ili trenutne situacije koje se dogode, događaju ili će se događati neko vrijeme. Također, u događaje se ubrajaju i stanja, tj. okolnosti u kojima je nešto istinito. U projektu *TimeML* uvode se i semantički razredi događaja, koji obuhvaćaju tradicionalne razrede iz prethodnih lingvističkih istraživanja, ali i neke nove razrede. *TimeML* razlikuje osam semantičkih razreda:

- OCCURRENCE – *umrijeti*, *sagraditi*;
- STATE – *voljeti*, *otet*;
- REPORTING – *reći*, *najaviti*;
- I_ACTION – *pokušati*, *obećati*;
- I_STATE – *vjerovati*, *namjeravati*;
- ASPECTUAL – *početi*, *završiti*;
- PERCEPTION – *vidjeti*, *čuti*.

Iz definicije događaja isključeni su ponavljajući događaji i generički događaji, tj. događaji koji opisuju svojstva neke opće skupine događaja, a ne pojedinog događaja.

Koristeći smjernice jezika *TimeML* izgrađen je korpus *TimeBank*, opisan u (Pustejovsky et al., 2003b). Cilj tog rada bila je izrada univerzalnog korpusa koji bi se mogao koristiti u većini lingvističkih istraživanja vezanih za ekstrakciju događaja i vremenskih relacija. Korpus sadrži 300 tekstova ručno označenih prema shemi *TimeML* i zadanim smjernicama. Tekstovi su odabrani kako bi pokrili područje medijskih izvora.

U nastavku slijedi opis nekoliko istraživanja koja koriste korpus *TimeBank*. Saurí et al. (2005) izradili su sustav *Evita* koji koristi statističke metode i zaključivanje temel-

jem pravila za označavanje događaja. Nakon početne predobrade, riječi i skupine riječi (engl. *chunk*) obrađuju se u dva koraka. U prvom koraku se određuje je li neka skupina riječi događaj, pri čemu se razmatraju samo skupine riječi koje su u predobradi označene kao imenice, glagoli ili pridjevi. U drugom se koraku određuju gramatičke značajke i dodatni atributi jezika *TimeML* poput vremena, aspekta i semantičkog razreda. Postignuta je vrijednost F_1-mjere od $80{,}12\%$.

U (Boguraev i Ando, 2005) provedena je kvantitativna analiza korpusa *TimeBank*. Pritom je taj korpus uspoređen s korpusima korištenim za neke druge zadatke na području obrade prirodnog jezika, poput korpusa *Penn Treebank*[1] korištenog za označavanje vrste riječi (engl. *part-of-speech tagging*). Malen broj riječi u korpusu određen je kao jedan od glavnih nedostataka korpusa. Također, detaljnijom analizom utvrđena je izrazito neravnomjerna distribucija semantičkih razreda događaja. Stoga su metodom profiliranja riječi (engl. *word profiling*) i na temelju različitih izračunatih statističkih vrijednosti pokušali ispraviti neke nedostatke Profiliranjem riječi su pritom iz frekvencija zajedničkoh pojavljivanja parova riječi kreirali značajke koje karakteriziraju pojedine riječi iz neoznačenog korpusa. Konačno, koristeći metodu minimizacije rizika (engl. *robust risk minimization*) izradili su linearni klasifikator kojim su na temelju lingvističkih značajki označavali nizove riječi kao događaje. Pritom je svakoj riječi bila dodijeljena jedna od tri oznake: *end* (zadnja riječ u nizu – pripada događaju), *inside* (riječ unutar niza – pripada događaju) i *outside* (riječ nije u nizu i nije dio događaja). Postignuta je vrijednost F_1-mjere od $64{,}0\%$ za označavanje događaja uz određivanje semantičkog razreda i $80{,}3\%$ samo za označavanje događaja. Profiliranje riječi pritom je donijelo poboljšanja od $2{,}7\%$ odnosno $1{,}7\%$.

Još jedan obećavajući postupak ekstrakcije događaja dan je u (Bethard i Martin, 2006). Slično kao i u (Boguraev i Ando, 2005), i ovdje su riječi klasificirane u tri razreda, samo što je umjesto razreda *end* ovdje uveden razred *begin*. Takav postupak također se naziva i BIO-dijeljenje teksta (engl. *BIO-chunking*) prema prvim slovima oznaka. Za klasifikaciju je odabrana metoda potpornih vektora (engl. *support vector machines*, SVM). Mjera točnosti metode dobivena je provođenjem peterostruke unakrsne provjere, pri čemu je bio korišten korpus *TimeBank*. Dobivena je vrijednost F_1-mjere od $75{,}9\%$ za označavanje događaja te $57{,}9\%$ za označavanje događaja i semantičkog razreda.

[1] http://www.cis.upenn.edu/~treebank/

2.2. Ekstrakcija vremenskih relacija

Ranija istraživanja na temu vremenskih relacija vezana su uz pitanje kako promatrati događaje u kontekstu vremena. Primjerice, u (C Bruce, 1972; Reichenbach, 1980) događaj predstavlja skup svih točaka između početne i završne točke. Vremenske relacije između dva tako definirana događaja određene su odnosima njihovih početnih i završnih točaka. Ovakvo razmatranje nije prikladno za definiranje svršenih predikata nad tim događajima. Naime, ako predikat vrijedi nad nekim skupom točaka, tada on vrijedi u svakoj točki tog skupa. Međutim, predikat *nacrtati krug* ne može vrijediti u svakoj točki skupa kojim je definirano trajanje pošto tek u završnoj točki vrijedi da je krug doista i nacrtan. Stoga je u (Allen, 1983) predložena intervalna logika koja događaje definira kao kontinuiranu cjelinu između početne i završne točke. Uz ovakvo razmatranje moguće je definirati predikat nad cijelim intervalom, čime se izbjegava problem iz prethodnog razmatranja. Relacije između intervala pritom su također određene odnosima početnih i završnih točaka intervala. Allenova intervalna definicija događaja doživjela je neke kritike i proširenja, primjerice u (Galton, 1990), gdje se proširuje uvođenjem točaka kao intervala nulte duljine.

Razvoj istraživanja vezanih uz ekstrakciju vremenskih relacija usko je vezan za istraživanja ekstrakcije događaja. Primjerice, Pustejovsky et al. (2003a) su pri izradi jezika *TimeML* uveli i oznake za vremenske relacije između dva događaja i između događaja i vremenskih izraza. Pritom je korištena Allenova intervalna logika te su uvedene oznake poput: *before*, *immediately before*, *includes*, *holds*, *simultaneous*, *identity*, *begins* i *ends*. Te oznake korištene su i u izradi korpusa *TimeBank* u (Pustejovsky et al., 2003b), no uz visoku razinu neslaganja označivača.

Uvidjevši da neslaganje označivača donosi velike poteškoće s korpusom *TimeBank*, Mani et al. (2006) pokušali su automatski pročistiti podatke. Smatrali su da je do neslaganja uglavnom došlo jer su različiti označivači promatrali različite parove događaja i vremena. Stoga su algoritmom vremenskog zatvorenja (engl. *temporal closure algorithm*), opisanim u (Verhagen, 2005), dodali vremenske relacije koje označivači nisu eksplicitno naveli, ali koje su se mogle zaključiti iz drugih, označenih relacija. Primjerice, ako je vrijedilo *A je prije B* i *B je prije C*, tada je zbog svojstva tranzitivnosti dodana relacija *A je prije C*. Zatim su naučili modele strojnog učenja da vremenskim relacijama daju odgovarajući tip: PRIJE, POSLIJE itd. Visoki rezultati (točnost $93{,}1\%$) bili su obećavajući, no interpretacija rezultata bila je otežana zbog algoritma vremenskog zaključivanja. Naime, iako je njime dobiven veći broj vremenskih relacija, nije bilo jasno odgovara li postignuta vjerojatnosna razdioba tih relacija stvarnoj situaciji.

Lapata i Lascarides (2004) odabrali su ponešto drukčiji pristup ekstrakciji vremenskih relacija. Odabrali su rečenice koje su sadržavale vremenske veznike poput *tijekom*, *prije*, *dok* itd. Zatim su jednostavnim probabilističkim modelom pokušali umetnuti ispravan vremenski veznik u rečenice čiji su vremenski veznici prethodno bili uklonjeni. Rezultati su bili obećavajući, uz postignutu točnost od $70{,}7\%$. Sličan pristup primijenili su i u (Lapata i Lascarides, 2006), gdje su dobili vrijednost F_1-mjere od $69{,}1\%$. Također, u okvirima tog rada odredili su preslikavanje korištenih vremenskih veznika na vremenske relacije definirane u jeziku *TimeML* te su dobiveni model testirali i na korpusu *TimeBank*. Pritom su postigli vrijednost F_1-mjere od 45.8%.

U cilju poticanja istraživanja metoda za određivanje vremenskih relacija organizirano je natjecanje *TempEval*, opisano u (Verhagen et al., 2007). Skupovi za učenje i ispitivanje sastojali su se od 162 dokumenta iz korpusa *TimeBank*, no vremenske relacije svedene su samo na oznake PRIJE, PREKLAPANJE i POSLIJE, i njihove disjunkcije. Pojednostavljenje je uvedeno s ciljem povećavanja slaganja označivača.

Tijekom natjecanja prijavljenim sustavima davali su se parovi događaja ili događaja i vremenskih izraza, a posao sustava bio je svakom paru dodijeliti odgovarajuću vremensku relaciju. Pritom su se u tri zadatka birali različiti parovi:

1. parovi događaja i vremenskih izraza unutar iste rečenice, primjerice u rečenici *Zrakoplovna tvrtka planira dnevne letove između Pariza i New Yorka* promatraju se događaji *planira* i *letove*;

2. najčešći događaji u dokumentu i vrijeme nastanka dokumenta, primjerice u rečenici *Predviđeno je lijepo vrijeme za iduća tri dana* koja se nalazi u dokumentu nastalom 25. svibnja 2011. godine upareni su događaj *predviđeno* i vrijeme *25. svibnja 2011. godine*;

3. parovi glavnih glagola u susjednim rečenicama, tj. glagola koji se nalaze najviše u sintaksnom stablu i služe kao glavni predikat rečenice; u rečenicama *Dizajn tornjeva nastao je prije 40 godina* i *Tornjevi su srušeni u napadu 11. srpnja* upareni su događaji *nastao* i *srušeni*.

U natjecanju su sudjelovali sustavi različitih arhitektura, no njihova točnost je uglavnom bila podjednaka. Kao referentna metoda uzeta je većinska metoda koja svim parovima u zadatku dodjeljuje najčešću oznaku za taj zadatak. Najviša točnost na prvom zadatku bila je 62%, na drugom 80%, a na trećem 55%. Pritom su postignuti tek marginalno bolji rezultati od referentne metode. Razina slaganja dosegla je vrijednost od 72% za kombinaciju prva dva zadatka i vrijednost od 65% za treći zadatak.

Bethard et al. (2007) je u svom radu odabrao nešto ograničeniji zadatak od onih iz natjecanja *TempEval*. Relacije su bile ograničene samo na događaje vezane sintaksnom strukturom glagol – rečenična dopuna glagolu (engl. *verb – clause pair*). Unatoč svojoj specifičnosti, ta se sintaksna struktura pokazala vrlo čestom u korpusu *TimeBank* – oko 20% događaja bilo je uključeno u takvu strukturu, a čak 50% susjednih parova glagolskih događaja činilo je upravo ovu strukturu. U radu su korištene različite leksičke, sintaksne i semantičke značajke te je metodom potpornih vektora (engl. *support vector machines*, SVM) postignuta točnost 89.2%. Ovo istraživanje pokazalo je da se metode strojnog učenja mogu dosta uspješno primijeniti na jednostavnije vremenske konstrukcije. Stoga bi se više specijaliziranih modela moglo povezati i iskoristiti za ekstrakciju složene vremenske strukture sadržane u tekstu.

Sva spomenuta istraživanja bavila su se tekstovima na engleskom jeziku. Cilj prvog dijela ovog rada jest istražiti mogućnosti ekstrakcije događaja iz tekstova na hrvatskom jeziku. Drugi dio rada bavit će se ekstrakcijom vremenskih relacija između označenih događaja. Koristit će se lingvističke značajke dobivene samo onim jezičnim alatima koji su dostupni za hrvatski jezik. U sklopu istraživanja izgradit će se i korpus na hrvatskom jeziku koji će pratiti neke smjernice dane pri izradi različitih korpusa na engleskom jeziku. Primjerice, koristit će se neki semantički razredi opisani u *TimeML*-u, ali i neki novi, ukoliko se uči potreba za njima. U eksperimentima će biti ispitana upotreba nekoliko različitih klasifikatora. U okviru vrednovanja točnosti ekstrakcije događaja provodit će se dva zadatka. U prvom zadatku odabrani klasifikatori koristit će se za pronalaženje događaja u rečenici. U drugom zadatku također će se pronalaziti događaji u rečenici, ali uz to će biti potrebno i odrediti semantički razred svakog događaja. Prilikom vrednovanja ekstrakcije vremenskih relacija promatrat će se točnost određivanja vrsta vremenskih relacija između različitih parova događaja.

3. Teorijske postavke

U ovom poglavlju objašnjene su teorijske postavke problema ekstrakcije događaja i vremenskih relacija. Dane su definicije događaja i vremenskih relacija te su definirani semantički razredi događaja i tipovi relacija koji će biti korišteni u istraživanju. Također je objašnjen odabrani pristup klasifikaciji. Konačno, zadnji dio daje kratke opise značajki prema kojima sustav određuje odgovarajući semantički razred ili vremensku relaciju, ovisno o zadatku.

3.1. Ekstrakcija događaja

3.1.1. Definicija događaja

Postoji nekoliko različitih definicija događaja. Prema (Pustejovsky et al., 2003a) događaj je pojam koji označava sve situacije koje se događaju. Pritom se uzimaju u obzir događaji koji su se ostvarili, trenutno traju ili će se ostvariti u budućnosti. Prema (Allan et al., 1998) događaj je neka jedinstvena stvar koja se dogodila u određenom trenutku u vremenu. Jedinstvenost događaja ograničava pojam događaja samo na konkretne instance nekih događaja. Primjerice, riječi koje opisuju neku klasu događaja preopćenite su, ne odnose se na jednu konkretnu instancu događaja te se zbog toga i ne smatraju događajima. Općenito, u svakom istraživanju potrebno je odrediti vlastitu definiciju događaja, ovisno o cilju istraživanja.

U okviru ovog istraživanja prihvaćene su definicije spomenute u prethodnom odlomku, ali uz neke preinake. Događaj obuhvaća sve događaje koji su se ostvarili, trenutno traju ili će se ostvariti u budućnosti. Ukoliko uz neku riječ stoji modalni glagol, ta riječ zbog promijenjene modalnosti neće se smatrati događajem. Također, stanja se neće smatrati događajima, ali promjene stanja hoće, pri čemu će takvi događaji dobiti posebnu oznaku. Riječi koje se smatraju događajima uvijek se moraju odnositi na neku konkretnu instancu događaja. Dozvoljeni su događaji koji se odnose na više instanci istog događaja ukoliko su te instance vezane širim kontekstom, no

takvi će se događaji posebno označavati. Riječi koje mogu opisivati događaj u ovom radu su imenice, glagoli i pridjevi.

3.1.2. Semantički razredi

Za razliku od radova poput (Bethard i Martin, 2006) i (Boguraev i Ando, 2005), opisanih u poglavlju 2, u ovom radu pri ekstrakciji događaja označava se samo jedna riječ te je za svaku riječ potrebno odrediti je li ona događaj ili nije. Dodatno, ako je riječ događaj, određuje se i njen semantički razred. Semantički razredi dijelom su preuzeti iz Pustejovsky et al. (2003a), no uvedeni su i neki novi radi boljeg slaganja s definicijom događaja danom u 3.1.1. Dozvoljeni su sljedeći semantički razredi:

- REPORTING – događaji u kojima je nešto objavljeno, deklarirano, informirano; ovakvi događaji imaju narativni karakter (*reći*, *kazati*, *objasniti*, *izjaviti*...);
- ASPECTUAL – događaji koji opisuju aspekt događaja, tj. je li neki događaj počeo, završio itd. (*započeti*, *okončati*, *završiti*...);
- PERCEPTION – događaji koji uključuju fizičku percepciju nekog drugog događaja (*vidjeti*, *ugledati*, *čuti*...);
- I_ACTION – događaji u ovom razredu predstavljaju akciju s namjerom, tj. događaj tipa I_ACTION otvara mjesto nekom drugom događaju koji mora biti eksplicitno naveden u tekstu *prihvatiti* ponudu, *izvršiti* napad...;
- OCCURRENCE – svi događaji koji opisuju da se nešto dogodilo (*happens* ili *occurs*); ovom razredu pripada najveći broj događaja;
- HALF_GENERIC – događaji koji do neke mjere imaju izražen karakter generičnosti, ali za koje je iz teksta jasno da se odnose na konkretnu radnju/radnje nekog subjekta; također skup od više konkretnih događaja (*utakmice* prvenstva...);
- STATE_CHANGE – događaji koji opisuju promjenu stanja nekog objekta ili osobe, pri čemu su uključene fizičke promjene, ali i promjene u psihičkim stanjima i razmišljanjima (*odlučiti*, *shvatiti*).

Detaljniji opisi semantičkih razreda popraćeni primjerima dani su u dodatku A koji sadrži upute za označavanje događaja.

Nekim događajima moguće je pridijeliti više semantičkih razreda. Primjerice, događaju *vidio* u nizu riječi ... *vidio je pad zrakoplova*... može se pridijeliti semantički

razred PERCEPTION, ali i semantički razred I_ACTION uz *pad* kao događaj kojemu događaj *vidio* otvara mjesto. Stoga su semantički razredi poredani po prioritetima:

1. PERCEPTION, ASPECTUAL, REPORTING;
2. I_ACTION;
3. OCCURRENCE, HALF_GENERIC, STATE_CHANGE.

Prva skupina semantičkih razreda najvišeg je prioriteta, dok je treća skupina najnižeg prioriteta. Prioriteti su određeni na temelju frekvencije pojavljivanja različitih semantičkih razreda u korpusu *TimeBank*, dane u (Boguraev i Ando, 2005). Odabirom ovih prioriteta razredi s manjim frekvencijama (poput razreda ASPECTUAL i PERCEPTION) neće se izgubiti odabirom drugog, češćeg razreda (npr. razred I_ACTION).

3.1.3. Značajke

Za ekstrakciju događaja korištene su različite lingvističke značajke. Zbog činjenice da su tekstovi na hrvatskom jeziku broj dostupnih jezičnih alata je ograničen. Zbog toga su korištene uglavnom leksičke značajke, iako postoje i neke sintaksne i semantičke značajke. Popis korištenih značajki za svaku riječi dan je u nastavku:

- riječ,
- lema,
- korijen riječi,
- vrsta riječi,
- padež,
- broj,
- modalitet,
- pomoćne riječi,
- razred u *Crovallexu*,
- glagolski način,
- negacija,
- okolne riječi.

Za svaku je riječ osim navedenih značajki moguće koristiti i sve značajke proizvoljnog broja prethodnih i sljedećih riječi. U nastavku je dan opis svake od navedenih značajki.

Riječ

Prva značajka je tekst same riječi koja se promatra. Zbog velikih slova na početku rečenice, svaka riječ ovdje je svedena samo na mala slova. Značajka se formira kao vreća riječi (engl. *bag-of-words*), tj. binarni vektor u kojemu je svakom elementu pridružena jedinstvena riječ iz skupa za učenje. Pri postavljanju značajke nekog primjera pronalazi se element čija je pridružena riječ jednaka tom primjeru i njegova vrijednost se postavlja na 1, dok su vrijednosti svih ostalih elemenata binarnog vektora jednake nuli.

Lema

Lema (engl. *lemma*) je osnovni (kanonski, natuknički) oblik riječi. Ukoliko je promatrana riječ imenica nekog oblika, njena lema bit će ta imenica u nominativu jednine. Ako je riječ glagol, tada je njena lema infinitiv tog glagola. Slično se može odrediti i za ostale vrste riječi. Za riječi sa identičnim imeničkim i glagolskim oblikom moguće je dobivanje većeg broja mogućih lema. Značajka je modelirana kao vreća riječi, ali za razliku od prethodne značajke, moguće je da više različitih elemenata binarnog vektora ima vrijednost jednaku 1.

Korijen riječi

Korijen je najmanji morfem koji ima neko značenje. Također se formira vrećom riječi. U ovom radu korištena je metoda S-1 opisana u (Šnajder, 2011). Ta metoda vrši korjenovanje riječi odbacivanjem sufiksa znakovnog niza do (uključivo) posljednjeg samoglasnika, pod uvjetom da je duljina uklonjenog sufiksa manja ili jednaka duljini ostatka (pseudokorijena). Iako jednostavan, ovaj postupak je opravdan činjenicom da većina obličnih nastavaka u hrvatskom jeziku započinje samoglasnikom.

Vrsta riječi

Vrsta riječi (engl. *part of speech*) je lingvistička kategorija riječi općenito definirana sintaksnim ili morfološkim ponašanjem riječi. Primjeri vrste riječi su imenica, glagol, pridjev itd. Za dobivanje vrste riječi u ovom radu korištene su oznake iz morfosintaktičkih opisa definiranih za hrvatski jezik prema normi MULTEXT-EAST u (Erjavec et al., 2003). Pritom je vrijednost ove značajke definirana prvim znakom morfosintaktičkog opisa.

Padež

Padež (engl. *case*) je gramatička kategorija flektivnog oblika kojim se izražava gramatička funkcija riječi. Dobiva se iz morfosintaktičkih opisa. Padeži hrvatskog jezika su: nominativ, genitiv, dativ, akuzativ, vokativ, lokativ i instrumental.

Broj

Broj (engl. *number*) se također dobiva iz morfosintaktičkih opisa. Moguće vrijednosti su *jednina* i *množina*. Značajka se dobiva iz morfosintaktičkog opisa riječi.

Modalitet

Modalitet (engl. *modality*) je značajka vezana samo za glagole. Označava pojavu modalnih glagola uz glavni glagol. Modalni glagoli pridodaju glavnom glagolu karakter sposobnosti, dozvole, želje ili obveze. U njih se ubrajaju glagoli *moći*, *trebati*, *morati*, *smjeti*, *željeti*, *htjeti* i *voljeti*. Modalnost glagola ispituje se jednostavnim pravilom: ukoliko je promatrana riječ glagol, tada se uzimaju u obzir najviše četiri prethodne riječi iz iste rečenice. Ako je neka od tih riječi modalni glagol i ako između modalnog glagola i promatrane riječi ne postoji neki drugi glagol, tada se modalnost promatrane riječi postavlja na vrijednost 1, inače ostaje postavljena na nulu. Prozor od četiri promatrane prethodne riječi procjenjen je razmatranjem dostupnih skupova i uočavanjem da su u tim tekstovima modalni glagol i njegov argument udaljeni za toliko riječi.

Pomoćne riječi

Pomoćni glagoli (engl. *auxiliary verbs*) su glagoli koji nadopunjuju sintaksno ili semantičko znanje o glavnom glagolu. U hrvatskom su to glagoli *biti* i *htjeti*. Za ostvarivanje značajke izrađen je popis svih oblika pomoćnih glagola. Ukoliko je neka riječ u okolini promatrane riječi pomoćni glagol, odgovarajućem elementu u binarnom vektoru pridružit će se vrijednost 1.

Sintaksno-semantički razred iz rječnika *Crovallex*

Crovallex je hrvatski valencijski rječnik glagola izrađen u sklopu istraživanja opisanog u (Preradovic et al., 2009). Sadrži 1739 glagola raspoređenih u 173 sintaksno-semantičkih razreda. Crovallex razred je sintaksno-semantički razred kojem neki glagol pri-

pada. Ovaj razred modeliran je kao binarni vektor čiji elementi označavaju pripadnost pojedinom sintaksno-semantičkom razredu.

Glagolski način

Glagolski način (engl. *grammatical mood*) je upotreba glagolskih nastavaka za izražavanje stava govornika prema onome što govore. U glagolske načine ubrajaju se *imperativ*, *infinitiv*, *kondicional*, *indikativ* i *particip*. Glagolski način dobiva se iz morfosintaktičkog opisa riječi.

Negacija

Ova značajka opisuje prisutnost negirajuće riječi u blizini promatrane riječi. U negirajuće riječi ubrajaju se sljedeće riječi: *ne*, *nisam*, *nisi*, *nije*, *nismo*, *niste*, *nisu*, *neću*, *nećeš*, *neće*, *nećemo* i *nećete*. Ukoliko se u okolini promatrane riječi nalazi neka od negirajućih riječi, ova se značajka postavlja na vrijednost 1.

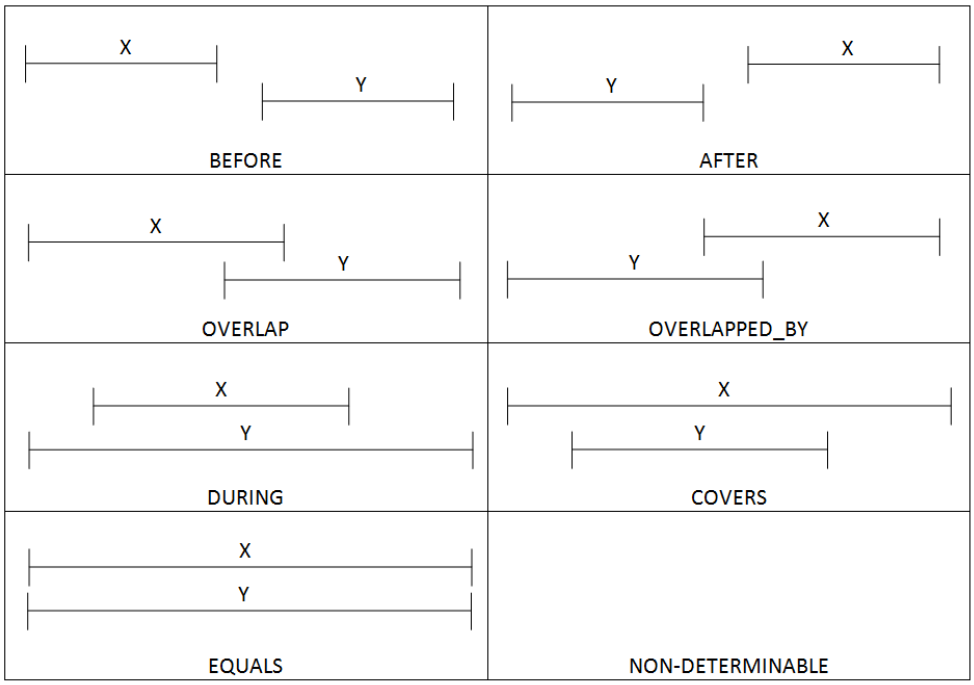
Okolne leme

Posljednja značajka korištena pri ekstrakciji događaja je skup lema u prozoru širine tri oko promatrane riječi. Ukoliko unutar prozora započinje nova rečenica, u ovaj skup ne ulaze leme iz susjedne rečenice. Značajka je modelirana kao vreća riječi.

3.2. Ekstrakcija vremenskih relacija

3.2.1. Definicija vremenske relacije

Vremenska relacija između dva događaja određena je odnosom početnih i završnih točaka događaja. U određivanju relacije uvijek se promatra prvi spomenuti događaj naspram drugog spomenutog događaja, tj. bitan je njihov redoslijed u tekstu. U sklopu ovog istraživanja događaji se razmatraju u okvirima proširene intervalne logike, definirane u (Allen, 1983; Galton, 1990). Događaji se prema tim definicijama smatraju kontinuiranom, cjelinom između početne i završne točke te je moguće definirati predikat nad cijelim intervalom. Dodatno, uvode se i točke kao intervali nulte duljine. Ovakva definicija omogućava dovoljnu fleksibilnost za opisivanje bilo kakvih vrsta događaja.



Slika 3.1: Tipovi vremenskih relacija.

3.2.2. Vrste vremenskih relacija

U sklopu ovog istraživanja korišteno je sedam vrsta relacija i posebna oznaka za neodredivu relaciju. U (Allen, 1983) definirane su sve moguće vremenske relacije između dva događaja, a u ovom radu radi pojednostavljivanja problema preuzete su samo neke. Relacija MEETS i njen inverz stopljeni su s relacijama BEFORE i AFTER, a relacije STARTS i FINISHES spojene su s postojećom relacijom DURING. Time Uz oznake *X* za prvi događaj i *Y* za drugi događaj, definicije korištenih vremenskih relacija dane su u nastavku:

- BEFORE (X before Y) – ovaj tip relacije obuhvaća sve relacije u kojima je prvi događaj u potpunosti prethodio drugom, tj. započeo je i završio prije početka drugog događaja. Iznimno, ovaj tip obuhvaća i one parove događaja gdje se trenutak završetka prvog događaja i trenutak početka drugog događaja podudaraju;
- AFTER (X after Y) – ovim tipom relacije obuhvaćene su one relacije u kojima prvi događaj slijedi nakon drugog događaja, tj. započinje nakon završetka drugog događaja. Također, ovaj tip obuhvaća i one parove događaja gdje se trenutak završetka drugog događaja i trenutak početka prvog događaja podudaraju. Ovaj tip relacije jednak je tipu BEFORE uz zamjenu uloga prvog i

drugog događaja;

- OVERLAP (X overlaps Y) – u ovaj tip ubrajaju se relacije u kojima je prvi događaj počeo prije početka drugog događaja, ali završio je nakon početka drugog događaja. Kod ovakvih relacija postoji određeni vremenski period u kojem su oba događaja trajala, tj. njihovo trajanje se djelomično preklapa;
- OVERLAPPED_BY (X overlapped by Y) – u ovaj tip relacije ubrajaju se relacije u kojima je drugi događaj počeo prije početka prvog događaja, ali završio je nakon početka prvog događaja. Kod ovakvih relacija postoji određeni vremenski period u kojem su oba događaja trajala, tj. njihovo trajanje se djelomično preklapa. Ovaj tip relacije odgovara tipu OVERLAP uz prethodnu zamjenu uloga prvog i drugog događaja;
- DURING (X during Y) – ovom tipu relacije pripadaju parovi događaja kod kojih je trajanje prvog događaja u potpunosti obuhvaćeno trajanjem drugog događaja. Drugim riječima, prvi događaj započeo je u trenutku započinjanja ili nakon početka drugog događaja te je završio prije završetka ili u trenutku završetka drugog događaja;
- COVERS (X covers Y) – ovom tipu relacije pripadaju parovi događaja kod kojih je trajanje drugog događaja u potpunosti obuhvaćeno trajanjem prvog događaja. Drugim riječima, drugi događaj započeo je u trenutku započinjanja ili nakon početka prvog događaja te je završio prije završetka ili u trenutku završetka prvog događaja. Ovaj tip relacije jednak je tipu DURING uz prethodnu zamjenu uloga prvog i drugog događaja;
- EQUALS (X equals Y) – ovaj tip relacije odnosi se na sve relacije čiji su događaji počeli i završili u jednakim trenutcima. Ovo je poseban slučaj tipova DURING i COVERS. Najčešće se ovom relacijom povezuju riječi koje se odnose na isti događaj;
- NON-DETERMINABLE – svaka dva događaja su u nekoj vremenskoj relaciji, tj. imaju nekakav odnos u vremenu. Međutim, iz teksta ponekad taj odnos nije očit. Zbog takvih slučajeva dodana je oznaka NON-DETERMINABLE. Ona se pridjeljuje svim parovima događaja čiji je odnos nemoguće odrediti iz teksta.

Ilustracija pojedinih relacija može se vidjeti na slici 3.1. Primjeri pojedinih vrsta relacija mogu se pronaći u dodatku B.

3.2.3. Značajke

Većina značajki koje se koriste u automatskoj ekstrakciji događaja koristi se i za ekstrakciju vremenskih relacija. U tu skupinu ubrajaju se sljedeće značajke: riječ, lema, korijen riječi, vrsta riječi, modalnost, pomoćne riječi i sintaksno-semantički razred. Te značajke definirane su za svaki događaj posebno. Dodatno, za određivanje relacije uvedena je i nova značajka koja modelira riječi koje se pojavljuju između dva događaja u tekstu.

Riječi između događaja

Ova značajka sadrži riječi koje se pojavljuju između dva događaja. Značajka je modelirana kao vreća riječi pri čemu se označuje samo prisutnost, a ne i frekvencija riječi.

4. Programska implementacija

U sklopu rada izrađen je skup različitih alata koji su služili kao programska podrška provođenju istraživanja. Za implementaciju tih alata odabran je programski jezik C# zbog mogućnosti koje donosi objektno orijentirana paradigma, a koje su pomogle u ostvarivanju generičke okoline primjenjive na zadatak ekstrakcije događaja, ali i na zadatak ekstrakcije vremenskih relacija. Izgrađeni alati mogu se podijeliti u nekoliko skupina. Prva skupina su alati i programski konstrukti potrebni za izgradnju odgovarajućeg modela domene i pripadnih entiteta (poput događaja i vremenskih relacija). Druga skupina su alati i konstrukti za podršku klasifikaciji, poput razreda za modeliranje različitih vrsta značajki, alata za ekstrakciju značajki itd. U tu skupinu pripada i posebna skupina jezičnih alata korištenih pri izgradnji značajki. Konačno, za dobivanje rezultata zaslužni su programi korišteni pri evaluaciji, a koji obuhvaćaju računanje različitih mjera i implementaciju korištenih evaluacijskih postupaka. Najvažniji razredi ovih skupina bit će ukratko opisani u nastavku.

Osim različitih alata spomenutih u prethodnom odlomku, u radu su se koristila i dva zasebna paketa slobodnog koda (engl. *open-source*). Prvi je programski paket *RapidMiner*,[1] sustav za analizu i dubinsko pretraživanje podataka. Drugi, *LibLinear*,[2] je biblioteka linearnih klasifikatora sposobnih za brzu obradu velikog broja primjera i značajki. Oba ova paketa također će biti opisana u nastavku.

4.1. Model domene

Domena ovog rada je ekstrakcija događaja i vremenskih relacija te model domene mora pružati podršku u modeliranju potrebnih entiteta: tokena (riječi i interpunkcijskih znakova), događaja, vremenskih relacija i dokumenata. Ovdje je dan kratak opis razreda kojima su ti entiteti realizirani i javnih metoda tih razreda.

Razred `MyToken` predstavlja tokene u sustavu, tj. sve riječi i interpunkcijske

[1] http://rapid-i.com/content/view/181/190/

[2] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

znakove koji se pojavljuju u tekstovima. Njegovi atributi su tekst koji modelira, ime dokumenta u kojem se pojavljuje i pozicija u tom dokumentu. Osim toga, nudi sljedeće metode za ispitivanje svojstava tokena:

- `bool IsPunctuation()` – ispituje je li token interpunkcijski znak;
- `bool IsWord()` – ispituje je li token riječ;
- `bool IsEndOfSentence()` – ispituje je li token završetak rečenice.

Ukoliko je neki token ujedno i događaj, on se modelira unutar razreda `Event`, gdje mu se pridružuju dodatna svojstva događaja. Ta dodatna svojstva uključuju semantički razred događaja te zastavice `Hypothetical` i `Uncertain` koje su služile kao pomoć pri označavanju. Dozvoljeni semantički razredi definirani su u zasebnom tipu `EventClass`.

Sljedeći važan entitet u modelu domene je vremenska relacija, realizirana razredom `Relation`. Atribute ovog razreda čine prvi i drugi događaj koje relacija povezuje te vrsta relacije. Slično kao i kod razreda, dozvoljene vrste relacije definirane su u zasebnom tipu `RelationClass`.

Posljednji entitet iz modela domene koji povezuje sve druge je `Document`. Podaci koje svaki objekt ovog razreda sadrži su ime datoteke, tekst, lista tokena, lista događaja i lista relacija. Javno sučelje dokumenta obuhvaća sljedeće metode:

- `EventClass GetEventClass(MyToken token)` – vraća oznaku semantičkog razreda predanog tokena ili oznaku `NOT_EVENT` ukoliko token nije događaj;
- `bool IsEvent(MyToken token)` – vraća `true` ako je predani token događaj, inače vraća false;
- `int GetTokenCount()` i `int GetWordCount()` – vraćaju broj tokena i riječi u dokumentu;
- `int GetTokenIndex()` – vraća indeks tokena, tj. njegov redni broj u listi svih tokena;
- `List<MyToken> GetWindow(MyToken token, int leftSize, int rightSize, bool breakOnPunctuation)` – vraća listu svih tokena koji se nalaze unutar zadanog broja riječi oko tokena. Ako je postavljena zastavica `breakOnPunctuation`, pretraživanje u određenu stranu staje ukoliko se pojavi token koji predstavlja interpunkcijski znak.

4.2. Programska podrška za klasifikaciju

Programska podrška za klasifikaciju je skupina alata i programskih objekata koja služi za pripremu podataka za klasifikaciju i njeno provođenje. U ovoj skupini najizraženija je upotreba principa objektnog oblikovanja kako bi ostvarila generička podrška i za događaje i za vremenske relacije, entitete modela domene koji se međusobno razlikuju po svojstvima. Stoga je u ovoj skupini potrebno istaknuti neka javna sučelja i apstraktne razrede te metode koje oni nude, dok će se konkretne implementacije tih sučelja i razreda tek spomenuti.

Apstraktni razred `IFeature` predstavlja osnovu bilo koje vrste značajki korištenih u sustavu. Zbog korištenja paketa *LibLinear* razred `IFeature` pohranjuje sve značajke u obliku binarnih vektora. Sukladno tome, nudi sljedeće metode:

- `abstract string GetName()` – vraća ime značajke;
- `int[] GetIndexes()` – vraća indekse svih elemenata binarnog vektora koji su postavljeni na `true`;
- `abstract int GetElementCount()` - vraća duljinu binarnog vektora, tj. broj elemenata.

Ovaj razred nasljeđuju svi razredi koji modeliraju značajke opisane u odjeljku 3.2.3: `AuxWords`, `BetweenWords`, `CrovallexClass`, `Lemma`, `Modal`, `Negation`, `NounAdjCase`, `Number`, `POS`, `Stem`, `SurroundingWords`, `VerbForm` i `Word`.

Za dobivanje nekih lingvističkih značajki u ovom radu koristi se morfološki leksikon, izrađen u sklopu (Šnajder, 2011), koji za danu riječ daje parove (*lema*, *morfosintaktički opis*). Morfosintaktički opisi za hrvatski jezik definirani su prema normi MULTEXT-East u (Erjavec et al., 2003) te sadrže vrijednosti pojedinih morfosintaktičkih kategorija sažeto kodiranih u jedan znakovni niz. Pošto spomenuti morfološki leksikon ne uzima u obzir kontekst riječi, moguće je dobivanje više parova (*lema*, *morfosintaktički opis*). Zbog takvih slučajeva značajke su modelirane kao binarni vektori. Pritom je svaka moguća vrijednost te značajke zaseban element binarnog vektora. U slučaju pojavljivanja neke vrijednosti značajke, vrijednost odgovarajućeg elementa binarnog vektora postavlja se na `true`, dok u suprotnom ostaje postavljena na `false`. Korištenje morfološkog leksikona omogućuje razred `Lemmatizer` sa sljedećim metodama:

- `List<string> GetLemmas(string word)` – vraća listu mogućih lema predane riječi;
- `public List<string> GetMSDTags(string word)` – vraća listu

morfosintaktičkih deskriptora predane riječi.

Implementacija S-1 korjenovanja opisanog u (Šnajder, 2011; Ljubešić et al., 2007) ostvarena je razredom `Stemmer`. Javno sučelje razreda sadrži sljedeće metode:

- `public string GetStem(string word)` – vraća korijen riječi dobiven S-1 korjenovanjem.

Tokenizacija teksta obavlja se uporabom razreda `Tokenizer`. Javne metode tog razreda su:

- `public List<MyToken> Tokenize(string text, string fileName)` – vraća listu svih tokena zadanog teksta;
- `public List<MyToken> TokenizeWords(string text, string fileName)` – vraća samo listu riječi zadanog teksta.

Sučelje `FeatureSet` modelira skup značajki koji se može pridijeliti nekom primjeru za učenje. Različite implementacije ovog sučelja mogu sadržavati različite vrste značajki. Upravo ovo sučelje predstavlja most između značajki događaja i značajki vremenskih relacija te nudi način kako u daljnjem postupku klasifikacije, uz upotrebu dodatnih sučelja spomenutih u nastavku, tretirati događaje i relacije kao da su jednaki. Uporaba bilo koje vrste značajki koja se koristi u ovom razredu može se omogućiti i onemogućiti, čime je stvoren mehanizam za odabir i uporabu bilo kojeg podskupa značajki. Ponuđene su sljedeće metode:

- `int[] GetIndexes()` – sve indekse svih značajki poredane uzlazno i skupljene tako da svi čine jedan binarni vektor koji predstavlja primjer za učenje;
- `int GetElementCount()` – vraća broj elemenata binarnog vektora;
- `void AddActiveType(object type)` – omogućuje korištenje predane vrste značajki;
- `void RemoveActiveType(object type)` – onemogućuje korištenje predane vrste značajki;
- `List<object> GetActiveFeatureTypes()` – vraća listu svih trenutno omogućenih vrsta značajki;
- `void ClearActiveFeatures()` – onemogućuje sve vrste značajki.

Ovo sučelje implementiraju razredi `EventFeatureSet`, `EventWindowFeatureSet` i `RelationFeatureSet`, koji modeliraju primjere za učenje odgovarajućih entiteta modela domene.

Izradu primjera za učenje na temelju predanog dokumenta i postavki omogućuje sučelje `IFeatureExtractor` i razredi `EventFeatureExtractor` i `RelationFeatureExtractor` koji ga implementiraju. Sučelje nudi sljedeće metode:

- `FeatureSet GetFeatureSet(object item, Document document, Hashtable parameters)` – vraća primjer za učenje iz zadanog dokumenta na temelju zadanih parametara;
- `object GetClassType(object item, Document document)` – vraća odgovarajući klasifikacijski razred, poput onih definiranih u `EventClass` i `RelationClass`, kojem pripada predani objekt;
- `List<object> GetAllFeatureTypes()` – vraća sve tipove značajki tog objekta;
- `List<object> GetActiveFeatureTypes()` – vraća sve omogućene tipove značajki tog objekta.

Klasifikatori korišteni u ovom sustavu predstavljeni su sučeljem `IClassifier`. Svi klasifikatori korišteni u ovom sustavu moraju implementirati sljedeće dvije metode:

- `void Train(string trainSetPath, string modelOutputPath)` – uči model primjerima za učenje pohranjenim na zadanoj putanji i ispisuje model na predviđenu putanju;
- `void Predict(string testInputPath, string modelInputPath, string resultOutputPath)` – učitava datoteku i model s predanih putanja te ispisuje rezultate na za to predviđeno mjesto.

Ovakav pristup korištenja klasifikatora predavanjem putanja do datoteka odabran je zbog potrebe za korištenjem paketa *LibLinear* koji radi upravo na taj način. U tu svrhu realiziran je razred `LibLinearWrapper` koji implementira sučelje `IClassifier` i koji se povezuje s paketom pohranjenim na datotečnom sustavu kako bi proveo treniranje i klasifikaciju. Osim toga, spomenuto sučelje koristi se i pri realizaciji referentnih klasifikatora, implementiranih razredima `EventBaseline` i `RelationBaseline`, dok su ostali klasifikatori implementirani u paketu *RapidMiner*.

4.3. Programska podrška za evaluaciju

Programska podrška za evaluaciju obuhvaća sve razrede koji se koriste za provođenje evaluacije i analizu rezultata. Ovdje će biti izdvojeni oni najvažniji, počevši s razre-

dom `EvaluationContext`. Ovaj razred objedinjuje sve informacije potrebne za provođenje evaluacije poput različitih odredišnih i izvorišnih putanja na datotečnom sustavu i zadanih parametara evaluacije. Također predstavlja most koji povezuje razrede iz ovog poglavlja s razredima iz prethodnog odjeljka. Stoga metode koje implementira, poput `void AddActiveFeature(object type)`, usmjeravaju pozive na druge, već opisane razrede te one neće ovdje biti opisivane.

Razred `CrossValidation` predstavlja servis koji implementira i provodi unakrsnu provjeru. Njegova jedina metoda je:

- `Results[] PerformCrossValidation(IClassifier classifier, List<List<EvaluationItem>> folds)` – provodi unakrsnu provjeru na zadanim, već podijeljenim primjerima uporabom predanog klasifikatora te vraća rezultate svih iteracija.

Razred `Results` služi kao spremnik vrijednosti mjera točnosti, dok je razred `EvaluationItem` samo pomoćni razred koji objedinjuje primjer (instancu razreda `FeatureSet`) i dokument.

Za odabir podskupa značajki koristi se razred `FeatureSubsetSelection` koji nudi dvije metode:

- `List<object> TopDownSelection(EvaluationContext context, List<List<EvaluationItem>> folds, string outputPath)` – vraća skup značajki koje su dale najbolje rezultate i ispisuje tijek rada na zadanu putanju; značajke se dobivaju *top-down* metodom;
- `List<object> BottomUpSelection(EvaluationContext context, List<List<EvaluationItem>> folds, string outputPath)` – vraća skup značajki koje su dale najbolje rezultate i ispisuje tijek rada na zadanu putanju; značajke se dobivaju *bottom-up* metodom.

4.4. Programski paket *RapidMiner*

RapidMiner[3] je okolina slobodnog koda za strojno učenje, dubinsku analizu podataka, prediktivnu analizu i poslovnu analizu. Koristi se u istraživanjima, edukaciji, razvoju prototipova i stvarnih aplikacija te u industrijske svrhe. S obzirom da je jedan od ciljeva ovog istraživanja istražiti mogu li se metode strojnog učenja koristiti za ekstrakciju događaja i vremenskih relacija, *RapidMiner* se pokazao kao vrlo dobar izbor za usporedbu mogućnosti različitih metoda i klasifikatora.

[3] http://rapid-i.com/content/view/181/190/

Rad u *RapidMineru* svodi se na manipulaciju različitim blokovima u radnom prostoru. Blokovi mogu predstavljati klasifikatore, transformacijske operatore, operatore za dohvat podataka, blokove za optimizaciju i evaluaciju itd. Povezivanjem blokova na odgovarajući način nastaje proces koji obavlja određenu zadaću. Za potrebe istraživanja u *RapidMineru* je izgrađen jednostavan proces koji se sastojao od repozitorija koji je predstavljao izvor podataka i bloka koji je provodio unakrsnu provjeru. Unutar tog bloka odabran je željeni klasifikator koji je na temelju podataka primljenih iz repozitorija vršio ekstrakciju događaja i vremenskih relacija. Prethodno je bilo potrebno pripremiti podatke u odgovarajućem formatu. Primjeri za učenje bili su zapisani u datoteku kao vrijednosti odvojene zarezom (engl. *comma-separated values*, CSV). Rezultati unakrsne provjere zapisivani su u odredišnu datoteku.

Zbog korištenja leksičkih značajki, poput riječi, lema i korijena riječi, uporaba *RapidMinera* bila je ograničena samo na određene klasifikatore. Naime, riječ je u *RapidMineru* predstavljena kao jedna značajka koja poprima određenu vrijednost ovisno o tome o kojoj se riječi radi. Svaka riječ pritom dobiva svoju jedinstvenu vrijednost. Međutim, dvije riječi mogu dobiti blisku vrijednost iako ni po čemu ne bi trebale biti bliske. Stoga pri klasifikaciji određenim metodama koje računaju udaljenost značajki može doći do pogreške. Klasifikatori na čiji rad ovakva dodjela oznaka ne utječe su Bayesov klasifikator i algoritam k najbližih susjeda te su se oni upotrijebili za ekstrakciju događaja i vremenskih relacija.

Drugi nedostatak *RapidMinera* je relativno nizak broj značajki s kojima može raditi. Pošto se leksičke značajke u ovom radu pretvaraju u binarne vektore, dimenzionalnost ulaznog prostora značajno raste i u tom slučaju *RapidMiner* se ne može koristiti.

4.5. Programski paket *LibLinear*

Alternativa *RapidMineru* koja može raditi s velikim brojem ulaznih primjera i značajki implementirana je u obliku programskog paketa *LibLinear*,[4] opisanog u (Fan et al., 2008). To je također paket slobodnog koda koji nudi mogućnost izrazito brze klasifikacije velikih količina podataka uporabom linearnih klasifikatora. U paketu su podržane logistička regresija i metoda potpornih vektora (engl. *support vector machines*, SVM) s linearnom jezgrom. Za uporabu *LibLineara* također je potrebno pripremiti podatke, no ovaj put u rijetkom obliku. Zbog toga su sve značajke bile binarizirane te su pohranjene u datoteci koja je bila predana *LibLinearu*. Ta je datoteka bila obliko-

[4] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

vana tako da je svaki primjer bio predstavljen jednim retkom u kojem je pisala oznaka razreda primjera i indeksi i vrijednosti samo onih elemenata binarnog vektora koji nisu bili jednaki nuli.

5. Eksperimentalno vrednovanje

Ovo poglavlje daje detaljan opis provedenog eksperimentalnog vrednovanja odabranih metoda strojnog učenja za ekstrakciju događaja i vremenskih relacija iz tekstova na hrvatskome jeziku. U nastavku je opisan pristup izgradnji i označavanju korpusa korištenog pri vrednovanju. Zatim slijedi opis provedenih eksperimenata za procjenu točnosti metoda strojnog učenja. Konačno, u zadnjem odjeljku prikazani su i komentirani dobiveni rezultati.

5.1. Izrada korpusa

Izrada korpusa obuhvaćala je tri koraka: odabir tekstova, označavanje događaja i označavanje vremenskih relacija. U prvom koraku odabrano je 230 novinskih članaka iz skupa članaka izašlih u *Vjesniku* u rasponu od 10 godina. Članci su birani s obzirom na svoju duljinu, vrstu i tematiku. Ciljana prosječna duljina bila je oko 500 tokena (riječi i interpunkcijskih znakova). S tematskog stajališta birani su članci koji su govorili o aktualnim događanjima u Hrvatskoj i svijetu, sportu, politici i kulturnim zbivanjima u Hrvatskoj. Članci poput kolumni, osvrta, kritika i drugih vrsta članaka koje se koriste za izražavanje subjektivnog dojma nisu bili uzimani u obzir. Konačni skup odabranih članaka sadrži 118.900 tokena, što obuhvaća 102.830 riječi s 26.095 različitih jedinstvenih oblika riječi i 10.963 jedinstvenih lema.

U drugom koraku izrade korpusa provedeno je označavanje događaja u člancima. Za potrebe ovog koraka napisane su upute za označavanje prikazane u dodatku A. Prateći te upute, četiri označivača označilo je prvi kalibracijski skup od deset članaka. Njihovo prosječno slaganje pri označavanju događaja, bez razmatranja odabranih semantičkih razreda, iznosilo je 0,5964. Ovo slaganje izraženo je prosjekom F_1-mjera između svih parova označivača zbog nesrazmjernog broja negativnih i pozitivnih slučajeva. Nakon prvog označavanja održan je kalibracijski sastanak čiji je cilj bio usuglašavanje označivača i razriješavanje razlika u oznakama. Na sastanku su uspješno razriješena tri dokumenta, a označivači su nakon sastanka uz novostečene spoznaje ponovo

Tablica 5.1: Rezultati označavanja događaja u korpusu.

Vrsta događaja	Broj pojavljivanja
OCCURRENCE	6867
REPORTING	1303
I_ACTION	1124
HALF_GENERIC	642
STATE_CHANGE	348
ASPECTUAL	301
PERCEPTION	58
ukupno	10643

označili ostalih sedam članaka iz prvog kalibracijskog skupa. Slaganje na tih sedam dokumenata dostiglo je vrijednost od 0,7676. Na drugom kalibracijskom sastanku razrješene su razlike u označavanju tih sedam dokumenata. Zatim je odabran novi kalibracijski skup od deset dokumenata na kojem je postignuto slaganje od 0,7951. Na temelju oznaka označivača i kalibracijskih sastanaka određene su konačne oznake za 20 članaka iz dva kalibracijska skupa. Ostalih 210 članaka bilo je ravnomjerno raspoređeno označivačima. Svaki označivač samostalno je označio događaje u dobivenim člancima, a dobivena razdioba semantičkih razreda dana je u tablici 5.1.

Treći korak izrade korpusa obuhvaćao je označavanje vremenskih relacija u tekstovima. Ovaj korak proveden je na sličan način kao i prethodni. Napisane su upute za označavanje, dane u dodatku B, te je provedeno označavanje prvog kalibracijskog skupa. Na sastanku su razriješene nedoumice u prvih šest članaka. Slaganje izraženo κ-statistikom, opisanom u (Cohen et al., 1960), na prvom kalibracijskom skupu iznosilo je 0,4861. Na drugom kalibracijskom skupu od deset članaka izmjereno slaganje bilo je jednako 0,5855. Zatim su označivači samostalno označili preostale članke. Dobivena razdioba tipova relacija dana je u tablici 5.2. Potrebno je istaknuti vrlo male vrijednosti pojavljivanja oznaka OVERLAP i OVERLAPPED_BY. Uzrok tome može biti rijetko pojavljivanje takvih vrsta relacija u novinskim tekstovima ili velika specifičnost tih vrsta relacija.

Tablica 5.2: Rezultati označavanja relacija u korpusu.

Vrsta relacije	Broj pojavljivanja
BEFORE	4860
AFTER	3500
EQUALS	1880
COVERS	1597
DURING	1341
NON-DETERMINABLE	763
OVERLAP	46
OVERLAPPED_BY	24
ukupno	14011

5.2. Eksperimenti

Za potrebe vrednovanja ekstrakcije događaja bila su provedena dva eksperimenta. U prvom eksperimentu sustav je za svaku riječ morao odrediti je li ona događaj ili nije. Zadatak u drugom eksperimentu bio je sličan prvom, uz razliku da je bilo potrebno odrediti i semantički razred svakog događaja. Provedeno je i eksperimentalno vrednovanje ekstrakcije vremenskih relacija, pri čemu je sustav trebao odrediti vrstu svake relacije. Svi eksperimenti provedeni su prema načelima opisanim u nastavku.

U prvom koraku vrednovanja korišteno je nekoliko različitih klasifikatora: naivan Bayesov klasifikator (engl. *naive Bayes classifier*), metoda k najbližih susjeda (engl. *k nearest neighbours*, k-NN) i metoda potpornih vektora (engl. *support vector machines*, SVM) s linearnom jezgrom. Pri vrednovanju prva dva klasifikatora korišten je programski paket *RapidMiner*, dok je pri procjeni točnosti metode potpornih vektora korišten paket slobodnog koda *LibLinear*. Prilagodbe značajki za uporabu u pojedinom paketu opisane su u poglavlju 4. Točnost svakog klasifikatora procijenjena je deseterostrukom unakrsnom provjerom. Provedeni su i eksperimenti uporabom odabranih referentnih klasifikatora (engl. *baseline*). Za zadatak ekstrakcije događaja referentni klasifikator je svakoj riječi dao njenu najčešću oznaku u korpusu na kojem je bio treniran, dok je u zadatku ekstrakcije vremenskih relacija svim relacijama bila dodijeljena najčešća oznaka korpusa. Klasifikator koji je dao najveću F_1^{macro}-mjeru odabran je za sljedeći korak vrednovanja.

U drugom koraku vrednovanja proveden je odabir podskupa značajki (engl. *feature subset selection*) koji je davao nabolje rezultate. Cilj ovog postupka bio je odabrati one

značajke koje najviše doprinose točnosti sustava. U slučaju ekstrakcije događaja korišten je postupak unaprijednog odabira značajki (engl. *forward feature selection*) u kojem se u svakoj iteraciji skup korištenih značajki proširuje onom značajkom koja najviše doprinosi točnosti sustava i koja poboljšava ukupnu točnost. Ovaj postupak je iterativne prirode i kreće od praznog skupa značajki. U prvoj iteraciji računa se točnost sustava korištenjem samo jedne od ponuđenih značajki. Nakon računanja promjene točnosti za svaku zasebno dodanu značajku, najbolja značajka dodaje se u skup. U idućem koraku svaka od preostalih značajki zasebno se dodaje dosad izgrađenom skupu, računa se nova točnost sustava i u skup se dodaje značajka s najvećim poboljšanjem. Postupak se prekida ako se iskoriste sve značajke ili ako u bilo kojem koraku niti jedna značajka ne uzrokuje poboljšanje točnosti sustava. Točnost sustava u ovom se koraku izražava F_1^{macro}-mjerom, a pri njenom računanju koristi se deseterostruka unakrsna provjera.

Pri odabiru značajki za ekstrakciju vremenskih relacija korišten je postupak unatražnog odabira značajki (engl. *backward feature selection*). Ovaj postupak korišten je zbog tehničkih razloga, prvenstveno zbog realizacije značajki koje dolaze u paru jer se računaju za svaku riječ posebno, poput padeža, vrste riječi itd. Za razliku od unaprijednog postupka, u unatražnom postupku početni skup sadrži sve značajke, a u svakoj iteraciji iz skupa se uklanja ona značajka čijim uklanjanjem se postiže najveće poboljšanje, ako takva značajka postoji. Postupak se prekida ako se izbace sve značajke ili ako izbacivanje bilo koje značajke donosi lošiji rezultat. Kao i u unaprijednoj metodi, točnost sustava i u ovom je koraku bila izražena F_1^{macro}-mjerom, za čije je računanje korištena deseterostruka unakrsna provjera.

Po završetku postupka odabira značajki dobiveni su i konačni rezultati za taj eksperiment, a koji su izraženi sljedećim mjerama: prosječna preciznost (engl. *precision*) za svaku klasu, prosječni odziv (engl. *recall*) za svaku klasu, prosječna F_1^{micro}-mjera i prosječna F_1^{macro}-mjera.

5.3. Rezultati

U ovom odjeljku prikazani su rezultati provedenih eksperimenta. Prvi dio odjeljka bavi se rezultatima vrednovanja ekstrakcije događaja, dok su u drugom dijelu prikazani rezultati vrednovanja ekstrakcije vremenskih relacija.

Tablica 5.3: Rezultati vrednovanja ekstrakcije događaja na različitim klasifikatorima.

	Dvije klase		Više klasa	
	$F_1^{macro}(\%)$	$F_1^{micro}(\%)$	$F_1^{macro}(\%)$	$F_1^{micro}(\%)$
Baseline	$47{,}28 \pm 0{,}48$	$79{,}91 \pm 1{,}06$	$24{,}93 \pm 4{,}94$	$74{,}21 \pm 1{,}68$
Bayes	$82{,}03 \pm 0{,}38$	$91{,}69 \pm 0{,}23$	$41{,}15 \pm 1{,}10$	$86{,}73 \pm 0{,}62$
3-NN	$83{,}98 \pm 0{,}73$	$94{,}37 \pm 0{,}29$	$50{,}27 \pm 2{,}57$	$92{,}56 \pm 0{,}24$
SVM	$\mathbf{88{,}44 \pm 3{,}17}$	$\mathbf{95{,}46 \pm 0{,}16}$	$\mathbf{54{,}21 \pm 2{,}80}$	$\mathbf{93{,}58 \pm 0{,}15}$

Tablica 5.4: Preciznost po klasama (%) - dvoklasni problem.

	Baseline	3-NN	Bayes	SVM
NOT_EVENT	$89{,}07 \pm 0{,}13$	$96{,}22 \pm 0{,}13$	$\mathbf{98{,}62 \pm 0{,}13}$	$97{,}16 \pm 0{,}13$
EVENT	$5{,}67 \pm 0{,}70$	$75{,}94 \pm 2{,}08$	$56{,}22 \pm 0{,}79$	$\mathbf{79{,}76 \pm 1{,}02}$

5.3.1. Ekstrakcija događaja

Rezultati prvog koraka eksperimenata prikazani su u tablicama 5.3 – 5.7. Tablica 5.3 prikazuje rezultate vrednovanja pojedinih metoda strojnog učenja, pri čemu su oni izraženi F_1^{macro} i F_1^{micro}-mjerama. Dobivene vrijednosti pokazuju da je određivanje semantičkog razreda događaja teži problem od određivanja je li neka riječ događaj. Postignute su visoke vrijednosti F_1^{micro}-mjera, no to je djelomično posljedica neuravnoteženog korpusa s mnogo negativnih primjera, tj. primjera koji nisu događaji. Štoviše, prema podacima iz odjeljka 5.1, samo 10% riječi u korpusu su ujedno i događaji. Veća razlika u rezultatima vidi se razmatranjem F_1^{macro}-mjere. Postignuta je najveća vrijednost F_1^{macro}-mjere od $88{,}44\%$ pri određivanju je li riječ događaj i $54{,}21\%$ pri određivanju semantičkog razreda riječi. Unatoč razlikama u točnosti, sve metode strojnog učenja u svim su zadacima bile bolje od referentne metode.

Tablice 5.4 i 5.5 prikazuju dobivene vrijednosti preciznosti i odziva po klasama

Tablica 5.5: Odziv po klasama (%) - dvoklasni problem.

	Baseline	3-NN	Bayes	SVM
NOT_EVENT	$88{,}43 \pm 1{,}22$	$97{,}55 \pm 0{,}26$	$92{,}01 \pm 0{,}29$	$\mathbf{97{,}80 \pm 0{,}14}$
EVENT	$6{,}01 \pm 0{,}80$	$66{,}81 \pm 1{,}20$	$\mathbf{88{,}84 \pm 1{,}07}$	$75{,}18 \pm 1{,}14$

Tablica 5.6: Preciznost po klasama (%) - višeklasni problem.

	Baseline	3-NN	Bayes	SVM
OCCURRENCE	$2{,}38 \pm 0{,}75$	$60{,}73 \pm 1{,}93$	$55{,}86 \pm 2{,}09$	$\mathbf{63{,}21 \pm 1{,}59}$
REPORTING	$0{,}63 \pm 0{,}79$	$81{,}21 \pm 2{,}68$	$67{,}57 \pm 2{,}33$	$\mathbf{81{,}45 \pm 3{,}00}$
ASPECTUAL	$0{,}48 \pm 1{,}51$	$\mathbf{65{,}48 \pm 8{,}37}$	$6{,}69 \pm 1{,}36$	$61{,}85 \pm 7{,}27$
PERCEPTION	$0{,}00 \pm 0{,}00$	$\mathbf{63{,}98 \pm 23{,}72}$	$14{,}87 \pm 5{,}99$	$63{,}20 \pm 24{,}46$
I_ACTION	$1{,}01 \pm 0{,}91$	$32{,}31 \pm 4{,}85$	$22{,}44 \pm 2{,}27$	$\mathbf{33{,}40 \pm 2{,}37}$
STATE_CHANGE	$0{,}87 \pm 1{,}15$	$23{,}45 \pm 7{,}67$	$16{,}64 \pm 2{,}59$	$\mathbf{37{,}02 \pm 9{,}65}$
HALF_GENERIC	$0{,}00 \pm 0{,}00$	$37{,}04 \pm 14{,}60$	$8{,}07 \pm 1{,}28$	$\mathbf{38{,}05 \pm 6{,}58}$
NOT_EVENT	$88{,}60 \pm 0{,}31$	$95{,}45 \pm 0{,}27$	$\mathbf{98{,}85 \pm 0{,}10}$	$96{,}78 \pm 0{,}19$

Tablica 5.7: Odziv po klasama (%) - višeklasni problem.

	Baseline	3-NN	Bayes	SVM
OCCURRENCE	$4{,}53 \pm 0{,}75$	$47{,}59 \pm 2{,}65$	$54{,}98 \pm 2{,}45$	$\mathbf{61{,}53 \pm 1{,}80}$
REPORTING	$0{,}77 \pm 0{,}96$	$72{,}46 \pm 5{,}88$	$\mathbf{85{,}12 \pm 2{,}59}$	$78{,}58 \pm 4{,}03$
ASPECTUAL	$0{,}33 \pm 1{,}05$	$53{,}47 \pm 9{,}09$	$\mathbf{74{,}76 \pm 5{,}41}$	$57{,}49 \pm 7{,}18$
PERCEPTION	$0{,}00 \pm 0{,}00$	$44{,}33 \pm 17{,}07$	$\mathbf{77{,}33 \pm 16{,}39}$	$53{,}33 \pm 17{,}84$
I_ACTION	$0{,}98 \pm 0{,}89$	$19{,}49 \pm 4{,}18$	$\mathbf{39{,}86 \pm 2{,}19}$	$19{,}13 \pm 2{,}38$
STATE_CHANGE	$0{,}62 \pm 0{,}80$	$23{,}24 \pm 8{,}89$	$\mathbf{51{,}38 \pm 6{,}73}$	$17{,}28 \pm 5{,}57$
HALF_GENERIC	$0{,}00 \pm 0{,}00$	$13{,}40 \pm 4{,}05$	$\mathbf{43{,}17 \pm 6{,}30}$	$26{,}44 \pm 6{,}17$
NOT_EVENT	$82{,}40 \pm 1{,}85$	$\mathbf{98{,}05 \pm 0{,}25}$	$90{,}17 \pm 0{,}64$	$98{,}01 \pm 0{,}14$

kad se određuje samo je li neka riječ događaj ili nije. Bayesov klasifikator pokazao je najveći odziv od svih klasifikatora, ali preciznost mu je slabija od 3-NN-a i SVM-a. Algoritam najbližih susjeda je demonstrirao relativno visoku preciznost, no slabiji odziv. Metoda potpornih vektora pokazala se otprilike jednako uspješnom i po pitanju preciznosti i po pitanju odziva.

Zadnje dvije tablice u ovom dijelu, 5.6 i 5.7, prikazuju dobivene vrijednosti preciznosti i odziva po klasama kad se određuje semantički razred događaja. Slično kao i u zadatku sa samo dvije klase, Bayesov klasifikator demonstrirao je visoki odziv, ali nešto nižu preciznost, dok je kod algoritma najbližih susjeda obrnut slučaj. Metoda potpornih vektora i u višeklasnom je problemu pokazala prosječno najbolje rezultate, što se očitovalo i u najboljim F_1-mjerama prikazanim u tablici 5.3. Promatranjem varijaciju preciznosti i odziva za pojedine klase, uočava se postojanje pravilnosti između

Tablica 5.8: Odabir podskupa značajki za ekstrakciju događaja i odgovarajuće vrijednosti F_1^{macro}-mjere (%).

Iteracija	1	2	3	4	5	6	7	8	9
riječ	47,78	52,91	53,51	55,32	55,71	55,80	55,81	55,95	55,94
lema	**52,02**	—	—	—	—	—	—	—	—
korijen riječi	49,29	52,04	54,30	55,02	55,33	55,48	55,48	55,84	55,62
vrsta riječi	66,66	52,51	**56,13**	—	—	—	—	—	—
padež	33,78	52,37	54,52	55,37	**57,24**	—	—	—	—
broj	66,66	52,28	54,49	55,40	57,05	**57,48**	—	—	—
modalnost	66,66	52,83	55,06	**56,75**	—	—	—	—	—
pomoćne riječi	63,34	**54,38**	—	—	—	—	—	—	—
Crovallex	37,55	52,43	54,33	55,05	56,75	57,07	57,26	**57,70**	—
glagolski način	54,51	54,20	55,00	55,45	56,32	56,69	56,51	56,89	56,80
negacija	66,66	53,11	54,65	55,31	57,12	57,31	**57,65**	—	—
okolne leme	20,25	53,89	53,63	55,21	55,61	54,72	54,71	55,31	55,51
odabrana značajka	lema	pomoćne riječi	vrsta riječi	modalnost	padež	broj	negacija	*Crovallex*	—

broja primjera određenog semantičkog razreda u korpusu, prikazanih u tablici 5.1. Semantički razredi koji su više zastupljeni u korpusu (poput OCCURRENCE i REPORTING) pokazali su manju varijaciju od onih manje zastupljenih (npr. PERCEPTION, ASPECTUAL).

Vrednovanje različitih klasifikatora pokazalo je da je metoda potpornih vektora najuspješnija od korištenih klasifikatora. Stoga je na njoj u drugom koraku proveden odabir podskupa značajki koje daju najbolje rezultate, tj. najveću vrijednost F_1^{macro}-mjere pri određivanju semantičkih razreda riječi. Primijenjena je metoda tipa "od dna prema vrhu" (engl. *bottom-up*) opisana u odjeljku 5.2, no na samom početku učinjena je iznimka. Naime, zbog neuravnoteženosti korpusa najveće rezultate na početku su davale značajke koje su zbog svoje jednostavnosti svim riječima pridjeljivale istu oznaku, NOT_EVENT. Zbog toga je u prvoj iteraciji odabrana lema kao prva značajka koja ulazi u skup. Odabir je izvršen na temelju prikazanih rezultata, po kojima je ta značajka bila pri samom vrhu, i činjenice da ti rezultati nisu bili posljedica neuravnoteženosti skupa, već ispravnog klasificiranja riječi. Postupak odabira podskupa značajki prikazan je u tablici 5.8. Svaki stupac predstavlja jednu iteraciju postupka, a za svaku neiskorištenu značajku prikazana je nova točnost klasifikacije ukoliko se ta značajka doda u izgrađeni skup. Ispod svakog stupca piše ime značajke koja je na kraju te iteracije odabrana i dodana u skup.

Konačno, rezultati deseterostruke unakrsne provjere nad odabranim značajkama prikazani su u tablici 5.9. Za zadatak određivanja semantičkog razreda riječi dobivene su sljedeće vrijednosti: $F_1^{macro} = (57{,}70 \pm 2{,}42)\%$, $F_1^{micro} = (76{,}96 \pm 0{,}65)\%$. Iz matrice zabune 5.9 vidljivo je da je preciznost po klasi općenito veća od odziva.

Bolji rezultati dobiveni su na semantičkim razredima OCCURRENCE, REPORTING, ASPECTUAL i PERCEPTION. To se može objasniti činjenicom da su razredi REPORTING, ASPECTUAL i PERCEPTION dosta usko definirani i specifični, pa su leksičke značajke dovoljne za uspješnu klasifikaciju. S druge strane, razredi I_ACTION, HALF_GENERIC i STATE_CHANGE daju dosta lošije rezultate jer su općenitiji i time bliski razredu OCCURRENCE. Toj općenitosti svjedoči i matrica zabune koja pokazuje da su svi ti razredi najčešće bili pogrešno proglašeni kao OCCURENCE ili uopće nisu bili označeni kao događaj. Lošiji rezultati razreda I_ACTION mogu se objasniti nedostatkom sintaksnih značajki jer sam razred puno više ovisi o kontekstu, tj. okolnim riječima i načinom na koji je s njima povezan.

		stvarno								
		OCCURRENCE	REPORTING	ASPECTUAL	PERCEPTION	I_ACTION	HALF_GENERIC	STATE_CHANGE	NOT_EVENT	Preciznost klase (%)
predviđeno	OCCURRENCE	**426,4**	11,8	3,7	0,3	47,3	30,4	10,8	17,7	77,78
	REPORTING	12	**107**	0,1	0,1	2,7	0,5	0,7	0,5	86,72
	ASPECTUAL	2,9	0,3	**17,3**	0	1,4	0,2	0,1	0,4	76,35
	PERCEPTION	0,7	0,2	0	**3,8**	0	0	0	0,1	83,92
	I_ACTION	21	0,6	0,7	0	**23**	0,7	1,2	1,6	47,93
	HALF_GENERIC	7,6	0,2	0,1	0	0,4	**7,2**	0	0,6	45,48
	STATE_CHANGE	6,1	0,8	1,1	0	3,3	0	**11,1**	1	48,79
	NOT_EVENT	210	9,4	7,1	1,6	34,3	25,2	10,9	**1042,4**	77,74
	Odziv klase (%)	61,08	85,38	53,33	80,00	25,00	14,06	29,41	97,46	

Tablica 5.9: Prosječna matrica zabune SVM klasifikatora s odabranim značajkama.

Tablica 5.10: Rezultati vrednovanja ekstrakcije vremenskih relacija na različitim klasifikatorima.

	$F_1^{macro}(\%)$	$F_1^{micro}(\%)$
Baseline	$34{,}69 \pm 0{,}05$	$6{,}44 \pm 0{,}00$
Bayes	$38{,}77 \pm 1{,}87$	$52{,}60 \pm 1{,}23$
3-NN (100 dok.)	$32{,}17 \pm 1{,}86$	$47{,}64 \pm 2{,}68$
SVM	$\mathbf{51{,}12 \pm 2{,}94}$	$\mathbf{64{,}16 \pm 1{,}07}$

5.3.2. Ekstrakcija vremenskih relacija

Rezultati vrednovanja ekstrakcije vremenskih relacija na različitim klasifikatorima dani su u tablicama 5.10, 5.11 i 5.12. U tablici 5.10 prikazane su F_1^{micro} i F_1^{macro}-mjere dobivene deseterostrukom unakrsnom provjerom. Odabrane metode strojnog učenja i u ovom su se eksperimentu pokazale boljim od referentne metode. Međutim, potrebno je napomenuti da, zbog velikih memorijskih zahtjeva i drugih tehničkih razloga, metodu najbližih susjeda nije bilo moguće provesti na svim dokumentima, već je ona provedena samo na slučajno odabranom podskupu od 100 dokumenata. Ova činjenica mora se uzeti u obzir pri daljnjim razmatranjima i usporedbama rezultata. Metoda potpornih vektora pokazala se kao najbolja i u ovom eksperimentu, postignuvši prosječnu vrijednost vrijednosti $F_1^{micro} = 64{,}16\%$ i $F_1^{macro} = 51{,}12\%$.

Tablica 5.11 prikazuje dobivenu preciznost po klasama u eksperimentu. Rezultati za referentni klasifikator su očekivani, jer je uz očuvanje razdiobe vrsta vremenskih relacija u svakom skupu unakrsne provjere najčešća oznaka bila BEFORE. Od ostalih vrsta relacija, najveću preciznost imaju BEFORE i AFTER, što se djelomično može pripisati i činjenici da su to najzastupljenije relacije. Najmanju preciznost imaju oznake OVERLAP i OVERLAPPED_BY. Zanimljivo je primijetiti da parovi BEFORE-AFTER i OVERLAP-OVERLAPPED_BY, koji predstavljaju inverzne parove relacija, imaju sličnu preciznost, dok to ne vrijedi za par DURING-COVERS u kojem oznaka DURING ima dosta veću preciznost od oznake COVERS. Sukladno prethodno prikazanim rezultatima, metoda potpornih vektora postiže najbolje rezultate.

U tablici 5.12 dane su dobivene vrijednosti odziva po klasama. Rezultati za referentni klasifikator pokazuju da je uvijek birana vrsta relacije BEFORE. Odzivi klasa OVERLAP i OVERLAPPED_BY pokazuju veliku varijaciju zbog malog broja primjera. Najveći odziv postigle su klase BEFORE, AFTER I DURING. Slično kao i u slučaju preciznosti, vidljiva je usklađenost inverznih parova relacija BEFORE-AFTER

Tablica 5.11: Preciznost po klasama (%).

	Baseline	3-NN (100 dok.)	Bayes	SVM
BEFORE	34,62 ± 0,05	60,89 ± 2,56	68,18 ± 1,18	**71,97 ± 2,12**
AFTER	0,00 ± 0,00	66,44 ± 5,20	61,66 ± 2,76	**70,77 ± 1,73**
OVERLAP	0,00 ± 0,00	0,00 ± 0,00	8,20 ± 8,73	**39,76 ± 28,77**
OVERLAPPED_BY	0,00 ± 0,00	0,00 ± 0,00	1,39 ± 1,26	**20,00 ± 24,28**
DURING	0,00 ± 0,00	51,78 ± 5,60	51,17 ± 3,47	**61,14 ± 4,28**
COVERS	0,00 ± 0,00	36,54 ± 5,82	33,89 ± 2,53	**52,95 ± 3,10**
EQUAL	0,00 ± 0,00	24,32 ± 1,79	41,54 ± 3,52	**45,74 ± 2,71**
NON-DETERMINABLE	0,00 ± 0,00	45,50 ± 8,23	39,69 ± 4,34	**58,13 ± 9,69**

Tablica 5.12: Odziv po klasama (%).

	Baseline	3-NN (100 dok.)	Bayes	SVM
BEFORE	100,00 ± 0,00	58,42 ± 4,50	59,34 ± 2,82	**74,36 ± 1,27**
AFTER	0,00 ± 0,00	48,77 ± 3,99	57,66 ± 2,63	**71,43 ± 2,05**
OVERLAP	0,00 ± 0,00	0,00 ± 0,00	**33,33 ± 26,06**	29,50 ± 17,07
OVERLAPPED_BY	0,00 ± 0,00	0,00 ± 0,00	**35,00 ± 47,43**	23,33 ± 26,29
DURING	0,00 ± 0,00	41,79 ± 4,13	**61,47 ± 5,42**	59,80 ± 2,48
COVERS	0,00 ± 0,00	18,58 ± 3,39	38,08 ± 2,99	**48,97 ± 4,43**
EQUAL	0,00 ± 0,00	**56,00 ± 4,73**	30,36 ± 4,97	46,33 ± 4,39
NON-DETERMINABLE	0,00 ± 0,00	31,61 ± 7,36	49,89 ± 8,76	**52,56 ± 7,48**

Tablica 5.13: Odabir podskupa značajki za ekstrakciju vremenskih relacija i odgovarajuće vrijednosti F_1^{macro}-mjere (%).

Iteracija	1	2	3	4
prvi – riječ	49,71	49,88	50,17	50,23
prvi – vrsta riječi	49,92	50,02	50,14	50,22
prvi – lema	49,94	**50,16**	—	—
prvi – korijen riječi	**50,01**	—	—	—
prvi – pomoćne riječi	49,95	50,06	50,16	50,24
prvi – modalnost	49,94	50,03	50,17	50,20
prvi – *Crovallex*	49,94	50,06	**50,20**	—
drugi – riječ	49,71	49,88	50,17	50,23
drugi – vrsta riječi	49,92	50,02	50,14	50,22
drugi – lema	49,94	50,16	49,87	49,96
drugi – korijen riječi	50,01	50,10	50,01	50,12
drugi – pomoćne riječi	49,95	50,06	50,16	50,24
drugi – modalnost	49,94	50,03	50,17	50,20
drugi – *Crovallex*	49,94	50,06	50,20	**50,42**
riječi između događaja	31,01	30,89	30,95	30,99
uklonjena značajka	prvi – korijen riječi	prvi – lema	prvi – *Crovallex*	drugi – *Crovallex*

i OVERLAP-OVERLAPPED_BY, dok je za klasu DURING postignut puno veći odziv od onoga za klasu COVERS. Metoda potpornih vektora postiže najbolje rezultate u najvećem dijelu klasa. Nešto viši odziv u nekim klasama postignut je uporabom naivnog Bayesovog klasifikatora.

U drugom koraku eksperimenta zbog postignutih rezultata korištena je metoda potpornih vektora. Proveden je metoda "od vrha prema dnu" (engl. *top-down*) opisana u odjeljku 5.2. Odabir je vršen na temelju rezultata prikazanih u tablici 5.13. Svaki stupac predstavlja jednu iteraciju postupka, a za svaku neiskorištenu značajku prikazana je nova točnost klasifikacije ukoliko se ta značajka doda u izgrađeni skup. Ispod svakog stupca piše ime značajke koja je na kraju te iteracije odabrana i dodana u skup.

		stvarno								
		BEFORE	AFTER	OVERLAP	OVERLAPPED_BY	DURING	COVERS	EQUALS	NON_DETERMINABLE	Preciznost klase (%)
predviđeno	BEFORE	**361,4**	39,2	1	0,1	14,8	31,1	41,3	13,8	71,97
	AFTER	37,9	**250**	0,5	0,5	13,2	19,8	23,6	7,9	70,77
	OVERLAP	0,4	0,4	**1,3**	0,1	0,6	0,2	0,7	0,1	39,76
	OVERLAPPED_BY	0,1	0,6	0	**0,6**	0,4	0,3	0,3	0	20,00
	DURING	13	12,6	0,7	0,4	**80,2**	5,8	15,8	3,1	61,14
	COVERS	27,7	15,8	0,3	0,3	5,9	**78,2**	15	4,5	52,95
	EQUALS	34,2	25	0,6	0,3	16,3	19,9	**67,1**	6,8	45,74
	NON_DETERMINABLE	11,3	6,4	0,2	0,1	2,3	4,4	4,2	**40,1**	58,13
	Odziv klase (%)	74,36	71,43	29,50	23,33	59,80	48,97	46,33	52,56	

Tablica 5.14: Prosječna atrica zabune SVM klasifikatora s odabranim značajkama.

Nakon odabira značajki dobiveni su sljedeći rezultati: $F_1^{macro} = (51{,}16 \pm 4{,}58)\%$, $F_1^{micro} = (64{,}60 \pm 1{,}62)\%$. U tablici 5.14 prikazana je matrica zabune metode potpornih vektora u kojoj su prikazane vrijednosti dobivene kao prosjeci deset iteracija unakrsne provjere. U većini slučajeva vidljivo je da je broj pogrešaka proporcionalan zastupljenosti pojedine klase u korpusu. Izuzetak čine inverzni parovi BEFORE-AFTER, OVERLAP-OVERLAPPED_BY i DURING-COVERS, gdje je postignut manji broj pogrešaka, no one su svejedno prisutne. Također, prosječan broj pogrešaka između parova klasa DURING-BEFORE, DURING-AFTER i COVERS-AFTER je nešto niži od, dok par COVERS-BEFORE ima veći broj pogrešaka, što može biti jedan od razloga niže preciznosti i odziva oznake COVERS od oznake DURING.

6. Zaključak

Danas su dostupne goleme količine pisanog teksta koje predstavljaju velik izvor znanja. Da bi se to znanje iskoristilo, potrebno je prepoznati različite entitete u tekstu te njihove uloge i odnose. Cilj ovog rada bio je ekstrakcija događaja i vremenskih relacija između događaja u pisanim tekstovima na hrvatskom jeziku. U tu svrhu proučena su srodna istraživanja i pristupi korišteni u njima. Određene su definicije entiteta, tj. događaja i relacija, karakteristične za ovo istraživanje. Provedeno je označavanje korpusa prikladnog za učenje i ispitivanje različitih pristupa ekstrakciji događaja i vremenskih relacija. Odabrane su značajke korištene u klasifikaciji, a koje su bile ograničene zbog nedostatka jezičnotehnoloških alata za hrvatski jezik. Implementirana je programska podrška za ekstrakciju značajki i klasifikaciju odabranim metodama strojnog učenja. Konačno, provedeno je eksperimentalno vrednovanje korištenih metoda, pri čemu su postignute vrijednosti F-mjera od 93% za označavanje događaja, 77% za određivanje semantičkog razreda događaja i 64% za određivanje vremenskih relacija.

Iako su postignuti obećavajući rezultati, nastavak istraživanja može se usmjeriti u različitim pravcima. S obzirom na složenost koncepta događaja i relativno nisku razinu slaganja označivača, u sklopu daljnjeg rada predlaže se detaljna analiza korištenog korpusa, njegovo proširivanje i razvoj metoda za automatsko ili poluautomatsko čišćenje korpusa. S druge strane, pojava novih jezičnih alata otvara vrata korištenju novih značajki. S obzirom da su u radu korištene pretežno leksičke značajke, proširenje istraživanja moglo bi se usmjeriti na izradu i iskorištavanje sintaksnih i semantičkih značajki koje bi davale puno više informacija o kontekstu. Konačno, povezivanje znanja o događajima i vremenskim relacijama i izrada baze znanja na temelju analiziranih dokumenata predstavljaju zanimljive izazove u budućnosti ovog istraživanja.

LITERATURA

J. Allan, J.G. Carbonell, G. Doddington, J. Yamron, i Y. Yang. Topic detection and tracking pilot study final report. 1998.

J.F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

S. Bethard i J.H. Martin. Identification of event mentions and their semantic class. U *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, stranice 146–154. Association for Computational Linguistics, 2006.

S. Bethard, J.H. Martin, i S. Klingenstein. Timelines from text: Identification of syntactic temporal relations. U *Semantic Computing, 2007. ICSC 2007. International Conference on*, stranice 11–18. IEEE, 2007.

S.J. Bethard. *Finding event, temporal and causal structure in text: A machine learning approach*. ProQuest, 2007.

B. Boguraev i R.K. Ando. Timebank-driven TimeML analysis. *Annotating, Extracting and Reasoning about Time and Events*, (05151), 2005.

B. C Bruce. A model for temporal references and its application in a question answering program. *Artificial intelligence*, 3:1–25, 1972.

J. Cohen et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

T. Erjavec, C. Krstev, V. Petkevič, K. Simov, M. Tadić, i D. Vitas. The MULTEXT-east morphosyntactic specifications for Slavic languages. U *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, stranice 25–32. Association for Computational Linguistics, 2003.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, i Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

A. Galton. A critical examination of Allen's theory of action and time. *Artificial Intelligence*, 42(2-3):159–188, 1990.

M. Lapata i A. Lascarides. Inferring sentence-internal temporal relations. U *Proceedings of HLT-NAACL*, stranice 153–160, 2004.

M. Lapata i A. Lascarides. Learning sentence-internal temporal relations. *Journal of Artificial Intelligence Research*, 27(1):85–117, 2006.

N. Ljubešić, D. Boras, i O. Kubelka. Retrieving information in Croatian: Building a simple and efficient rule-based stemmer. *Digital information and heritage/Seljan, Sanja*, stranice 313–320, 2007.

I. Mani, M. Verhagen, B. Wellner, C.M. Lee, i J. Pustejovsky. Machine learning of temporal relations. U *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, stranice 753–760. Association for Computational Linguistics, 2006.

N.M. Preradovic, D. Boras, i S. Kisicek. CROVALLEX: Croatian verb valence lexicon. U *Information Technology Interfaces, 2009. ITI'09. Proceedings of the ITI 2009 31st International Conference on*, stranice 533–538. IEEE, 2009.

J. Pustejovsky. The syntax of event structure. *Cognition*, 41(1):47–81, 1991.

J. Pustejovsky, J. Castano, R. Ingria, R. Sauri, R. Gaizauskas, A. Setzer, G. Katz, i D. Radev. TimeML: Robust specification of event and temporal expressions in text. *New Directions in Question Answering*, 2003:28–34, 2003a.

J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, et al. The TimeBank corpus. U *Corpus Linguistics*, svezak 2003, stranica 40, 2003b.

H. Reichenbach. *Elements of symbolic logic*. Dover Publications, 1980.

R. Saurí, R. Knippen, M. Verhagen, i J. Pustejovsky. Evita: a robust event recognizer for QA systems. U *Proceedings of the conference on Human Language Technology*

and Empirical Methods in Natural Language Processing, stranice 700–707. Association for Computational Linguistics, 2005.

E.V. Siegel i K.R. McKeown. Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–628, 2000.

J. Šnajder. Morfološka normalizacija tekstova na hrvatskome za dubinsku analizu i pretraživanje informacija. 2011.

Z. Vendler. Verbs and times. *The philosophical review*, 66(2):143–160, 1957.

M. Verhagen. Temporal closure in an annotation environment. *Language Resources and Evaluation*, 39(2):211–241, 2005.

M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, i J. Pustejovsky. Semeval-2007 task 15: Tempeval temporal relation identification. U *Proceedings of the 4th International Workshop on Semantic Evaluations*, stranice 75–80. Association for Computational Linguistics, 2007.

H. Verkuyl. Aspectual composition: Surveying the ingredients. *Perspectives on aspect*, stranice 19–39, 2005.

Dodatak A
Upute za označavanje događaja

Obrada prirodnog govora područje je čije su praktične primjene moguće u različitim dijelovima ljudskog života. Da bi računalo saznalo određenu informaciju iz nekog izvora prirodnog govora, primjerice pisanog teksta, često je potrebno obaviti mnoge operacije poput analize teksta, generiranja strukturiranih podataka itd. Jedna od mogućih primjena obrade prirodnog govora je analiza teksta u svrhu otkrivanja događaja i vremenskih relacija između događaja u tekstu te izrada strukturiranih podataka koji sadrže znanje o kronološkom uređenju tih događaja.

U ovom je zadatku potrebno označiti događaje u novinskim tekstovima na hrvatskom jeziku. Na temelju označenih događaja u sljedećem zadatku označavat će se vremenske relacije između tih njih, no taj zadatak rješavat će se u kasnijoj fazi. U nastavku dokumenta dan je opis zadatka označavanja događaja. Taj opis obuhvaća objašnjenje pojma događaj u kontekstu zadatka, neka pravila/konvencije vezane uz jezične strukture koje olakšavaju označavanje te primjeri koji prikazuju što se treba, a što ne treba označavati.

A.1. Što je događaj?

U mnogim istraživanjima dane su različite definicije pojma događaj, ovisno o konkretnom zadatku obrađivanom u istraživanju. U kontekstu ovog označavanja događajem se smatra svaka radnja ili situacija koja se obavila, dogodila, pojavila, traje itd., tj. svaka akcija ili djelovanje, neovisno o svom vremenskom trajanju. Događaji mogu biti trenutačni ili mogu trajati kroz period, mogu biti jednostavni ili složeni (sastoje se od puno manjih događaja), mogu biti realni (dogodili su se ili se događaju u stvarnom svijetu) ili hipotetski (nisu se dogodili, ali o njima se priča u nekom hipotetskom kontekstu, ili će se možda tek dogoditi). Događajima se ne smatraju predikati koji opisuju stanja ili okolnosti u kojima nešto vrijedi, kao ni mentalna stanja, želje, namjere i sl. Također,

događajima se u kontekstu ovog istraživanja neće smatrati predikati koji označavaju generičke radnje, tj. radnje koje opisuju općeniti koncept ili neodređeni skup događaja istog tipa.

U nastavku slijede primjeri rečenica s riječima koje treba ili ne treba označiti kao događaje. One riječi koje je potrebno označiti bit će **podebljane i podcrtane**, a riječi koje su potencijalni kandidati za označavanje, ali ih se u ovom kontekstu ne treba označiti, bit će **podebljane**. Dodatno, ukoliko se u određenom dijelu teksta opisuje specifična radnja, vrsta riječi, koncept itd., riječi na koje se taj dio teksta odnosi bit će napisane crvenom bojom.

A.2. Što označavati?

Kao uvod u daljnja objašnjenja, u nastavku su prikazani neki primjeri događaja koje je potrebno označiti. Ukoliko se događaj sastoji od više riječi, označena je samo ona riječ koja nosi najveći dio značenja (nositelj događaja). Događaji su vrlo često izraženi glagolima, no mogu biti označeni i drugim vrstama riječi, poput imenica, pridjeva itd. Uz primjere su dani i razlozi zašto je nešto označeno kao događaj.

*Za potrebe Vojnih **igara** 2012. godine u studentske domove bit će uloženo 50 milijuna kuna.*

- događaj *igre* označen je jer predstavlja određenu instancu Vojnih igara (one Vojne igre koje će se održati u 2012. godini);
- događaj *uloženo* označava događaj koji je već počeo ili koji će tek početi, a koji će se završiti negdje u budućnosti.

*Vjesnikova književna nagrada »Ivan Goran Kovačić« za godinu 1998. **dodijeljena** je Zvonimiru Berkoviću za knjigu »Dvojni portreti«*

- događaj *dodijeljena* označava neki događaj koji se dogodio u prošlosti.

*Nino Bule, najbolji nogometaš Zagreba, unatoč nekim inozemnim **ponudama**, po svemu će sudeći ostati u Kranjčevićevoj.*

– riječ *ponudama* označena iako označava skup više događaja jer je taj skup točno određen (odnosi se točno na one ponude koje su upućene nogometašu).

A.3. Što ne označavati?

1. Izrazi koji iskazuju stanja, mišljenja, razmatranja, želje, stavove itd. ne označavaju se kao događaji.

   *Tajnik Mjesnog odbora u Mirkovcima Mile Madžar **procjenjuje** da su Hrvati iz BiH **kupili** 70 srpskih kuća, što **potvrđuje** značajnu **promjenu** demografske strukture stanovništva na tom području.*

   *Akademik Katičić je međunarodno **priznati** stručnjak na području sintakse i povijesti hrvatskoga književnog jezika, teorije genetske lingvistike...*

   *Given je, naime, **zaprepašten** ping-pongom oko **utakmice** u Dublinu.*

   *Vratar australske reprezentacije i engleske Aston Ville, Mark Bosnich **vjeruje** kako će uskoro **potpisati** ugovor s aktualnim europskim prvakom, Manchester Unitedom, nakon što **obavi** sve potrebne **razgovore** i liječnički **pregled**.*

   ***Činimo** sve što možemo, ali svakako bismo **htjeli** još i više, kako bi naš Grad **postao** prava kolijevka kulture.*

2. Generički izrazi koji označavaju općeniti skup događaja istog tipa i koji se ne odnose na bilo kakve konkretne (stvarne ili hipotetske) instance tih događaja također se ne označavaju.

   *Grad Kaštela **kandidirat** će se za domaćina Svjetske **izložbe** »Expo 2004«., što se svake druge godine **održava** u jednom svjetskom gradu i na kojoj **sudjeluju** zemlje koje imaju izlaz na mora ili oceane.*

Praizveden *u Švedskoj početkom ove godine, a potom* ***izvođen*** *u Norveškoj te u amsterdamskom kazalištu »Frescati«, scenski projekt »Fragile«* ***ispituje*** *opstojnost mita u našoj istrošenoj stvarnosti.*

U tim dokumentima ne mogu se ***naći*** *imena i odgovornost pojedinca.*

Kao da ***slušamo*** *staru ploču ili pokvareni gramofon.*

3. Fraze koje se sastoje od modalnog glagola (*trebati*, *moći*, *smjeti*, *znati*, *htjeti*, *morati*, *željeti*, *voljeti*) i glavnog glagola ne označavaju događaj, već se tretiraju kao stanje.

   No kulturne ustanove ne ***moraju*** *se* ***oslanjati*** *samo na gradski proračun, nego* ***mogu*** *svojim djelatnicima* ***povećati*** *plaće i iz vlastitih prihoda.*

A.4. Kako označavati?

Događaji u tekstu uvijek se označavaju samo jednom riječju, koja se naziva nositeljem događaja.

1. Ako je događaj izražen kao glagolska fraza (*kandidirat će se*, *dodijeljena je*) koja se sastoji od glavnog glagola i pomoćnih glagola, potrebno je označiti samo glavni glagol, tj. riječ koja nosi tu glagolsku frazu.

   Grad Kaštela ***kandidirat*** *će se za domaćina…*

   Vjesnikova književna nagrada »Ivan Goran Kovačić« za godinu 1998. ***dodijeljena*** *je…*

2. Ako je događaj izražen glagolskom frazom koju čine aspektualni glagol (*započeti*, *završiti*, *okončati*…) i glavni glagol, potrebno je označiti oba glagola. Pritom svaki događaj dobiva posebnu oznaku (objašnjeno u nastavku). Ta dva glagola zapravo predstavljaju dva dijela istog događaja: jedan je npr. početak, a

drugi je cijeli događaj.

*Visoke temperature koje su nas posljednjih dana – ipak **iznenadile**, već su **počele uzrokovati** zdravstvene probleme.*

3. Ako se radi o imeničkoj frazi (npr. *Vojne igre*), tada je kao događaj potrebno označiti samo glavnu riječ fraze (*igre*), a ne cijelu frazu (riječ *Vojne*).

   *Za potrebe Vojnih **igara** 2012. godine u studentske domove...*

4. Ako je događaj naveden imeničkom frazom popraćen s glagolskim predikatom, potrebno je označiti oboje kao događaje. Svaki događaj dobiva posebnu oznaku (objašnjeno u nastavku).

   ***Pregovori** su **održani** prošle srijede.*

   *Sutrašnja **utakmica privući** će mnoge obožavatelje željne dugo iščekivanog **okršaja** između dvaju najjačih klubova lige.*

A.5. Vrste događaja

Svakom događaju potrebno je dodijeliti vrstu, odnosno razred kojem pripada. Na temelju različitih semantičkih svojstava u okviru ovog označavanja događaji se mogu podijeliti u sljedeće razrede:

- OCCURRENCE,
- PERCEPTION,
- REPORTING,
- ASPECTUAL,
- I_ACTION,
- HALF_GENERIC,
- STATE_CHANGE.

Dodatno, osim razreda događaja, za svaki događaj potrebno je odrediti je li on realan ili hipotetski. Slijedi detaljniji opis događaja koji pripadaju pojedinom razredu.

1. REPORTING – U ovu kategoriju spadaju događaji koji opisuju akcije u kojima osobe ili organizacije nešto objavljuju, deklariraju ili informiraju. Ovakvi događaji imaju narativni karakter. Primjeri glagola koji imaju narativni karakter jesu: *reći*, *prijaviti*, *kazati*, *objasniti*, *izjaviti*...

   Među 300 važnih sredozemnih šumskih područja nalaze se i hrvatski otoci Elafiti (kod Dubrovnika), Lastovnjaci, otok Plavnik, dva predjela na Biokovu, Učka i donji tok Neretve, ***izvijestio*** *je u srijedu Pokret prijatelja prirode »Lijepa naša«.*

   »Černomirdin je potom ***govorio*** *o pojedinostima plana«,* ***rekao*** *je Sergejev, ali nije* ***naveo*** *kako je reagirao Milošević.*

   Kurdski pobunjenici ***ubili*** *su tri turska vojnika u bombaškom* ***napadu*** ***izvršenom*** *u ponedjeljak nakon što je* ***počelo*** ***suđenje*** *kurdskom vođi Abdullahu Öcalanu, kojemu prijeti smrtna kazna zbog optužbe za* ***izdaju***, ***priopćili*** *su u utorak turski vojni izvori.*

2. PERCEPTION – U ovaj razred spadaju događaji koji uključuju fizičku percepciju nekog drugog događaja. Takvi događaji često su opisani glagolima poput: *vidjeti*, *ugledati*, *čuti* itd.

   S druge strane, u ***procesu*** *protiv Roberta Faurissona, koji je* ***nijekao*** *postojanje plinskih komora, francuski suci* ***došli*** *su u Poljsku* ***vidjeti*** *ih na svoje oči.*

   Svjedoci su ***rekli*** *policiji da su* ***čuli*** ***pucnjeve*** *u noći.*

3. I_ACTION – Događaji ovog razreda predstavljaju "akciju s namjerom" (engl. *intentional action*). Događaj koji pripada ovom razredu otvara mjesto drugom događaju koji mora biti eksplicitno naveden u tekstu. Taj drugi događaj opisuje akciju ili situaciju iz koje nešto možemo zaključiti na temelju njegove relacije s događajem koji je tipa I_ACTION. Ovaj razred često se javlja u strukturi glagol

– direktni objekt (koji je također događaj), pri čemu se glagol označava oznakom I_ACTION.

*U povodu Dana državnosti, **misu** za domovinu u crkvi Svetoga Marka **predvodio** je u subotu nadbiskup zagrebački msgr. Josip Bozanić.*

***Pokazao** je kako u današnjem teatru ne **postoje** rubne ili središnje uloge.*

*Anto Đapić, predsjednik Hrvatske stranke prava, **pozvao** je Hrvatsku kršćansku demokratsku uniju Marka Veselice da zajedno **nastupe** na sljedećim **izborima**.*

*Šeparović je **izjavio** da u hrvatskom pravosuđu ima mnogo problema, ali da se sada **poduzimaju** **mjere** kako bi se **ubrzao**, **pojednostavio** i **učinio** konstruktivnim cijeli **postupak** pred sudom.*

*Kurdski pobunjenici **ubili** su tri turska vojnika u **bombaškom** napadu **izvršenom** u ponedjeljak nakon što je **počelo** **suđenje** kurdskom vođi Abdullahu Öcalanu, kojemu prijeti smrtna kazna zbog optužbe za **izdaju**, **priopćili** su u utorak turski vojni izvori.*

Moguće je i ulančavanje više događaja u lanac: svi događaji osim zadnjeg u tom slučaju trebaju biti označeni oznakom I_ACTION, primjerice:

*Izrael je **zamolio** SAD da **odgodi** vojni **napad** na Irak.*

Ukoliko se u tekstu javlja par događaja koji su neposredno jedan iza drugog, a koji se odnose na isti događaj, pri čemu je jedan imeničkog, a drugi glagolskog karaktera, potrebno je označiti oboje. Pritom se glagolski događaj označava kao I_ACTION.

***Prosvjed** je **održan** na glavnom trgu.*

4. ASPECTUAL – U ovaj razred spadaju događaji koji opisuju aspekt događaja, tj. je li neki događaj počeo, završio itd.

*Time je **proces** iskapanja na tom lokalitetu **okončan**.*

***Pripreme** za novu sezonu Bilić će **započeti** u lipnju u Splitu, a ovih dana je na odmoru u Monte Carlu.*

*Kurdski pobunjenici **ubili** su tri turska vojnika u bombaškom **napadu** **izvršenom** u ponedjeljak nakon što je **počelo** **suđenje** kurdskom vođi Abdullahu Öcalanu, kojemu prijeti smrtna kazna zbog optužbe za **izdaju**, **priopćili** su u utorak turski vojni izvori.*

5. OCCURRENCE – Najveći broj događaja spada u ovaj razred. Ovaj razred obuhvaća sve događaje koji opisuju da se nešto dogodilo (engl. *happens, occurs*).

   *Za potrebe Vojnih **igara** 2012. godine u studentske domove bit će **uloženo** 50 milijuna kuna.*

   *»Černomirdin je potom **govorio** o pojedinostima plana«, **rekao** je Sergejev, ali nije **naveo** kako je **reagirao** Milošević.*

   *Anto Đapić, predsjednik Hrvatske stranke prava, **pozvao** je Hrvatsku kršćansku demokratsku uniju Marka Veselice da zajedno **nastupe** na sljedećim **izborima**.*

   *Ruski ministar vanjskih poslova Igor Ivanov **otputovao** je u utorak u trodnevni **posjet** Kini, **izvijestila** je agencija Itar-Tass.*

   *Kurdski pobunjenici **ubili** su tri turska vojnika u bombaškom **napadu** **izvršenom** u ponedjeljak nakon što je **počelo** **suđenje** kurdskom vođi Abdullahu Öcalanu, kojemu prijeti smrtna kazna zbog optužbe za izdaju, **priopćili** su u utorak turski vojni izvori.*

6. HALF_GENERIC – u ovaj razred spadaju događaji koji imaju do neke mjere izražen karakter generičnosti, ali za koje je iz teksta jasno da se odnose na

konkretnu radnju/radnje nekog subjekta. Događaji u ovom razredu često mogu podrazumijevati više manjih stvarnih događaja, ali se u kontekstu spominjanja doživljavaju kao jedna aktivnost koja grupira te manje događaje.

*Eto, nakon Maria Stanića na popisu su Robert Kovač, naše veliko otkriće u **utakmicama** s Italijom i Španjolskom, a po najnovijim vijestima moglo bi biti i problema s **nastupom** Daria Šimića.*

*Osobno sam **upućivao** **pozive** da **ostanu**, no oni su **slijedili** zapovijedi čelnika iz Srbije i **otišli**.*

7. STATE_CHANGE – u ovaj razred spadaju događaji koji označavaju da je došlo do promjene stanja nekog objekta ili osobe. Promjene stanja uključuju promjene fizičkih stanja, ali i promjene u psihičkim stanjima i razmišljanjima.

*Istarski kratkodlaki i istarski oštrodlaki gonič konačno su **priznati** kao autohtone hrvatske (lovne) pasmine pasa.*

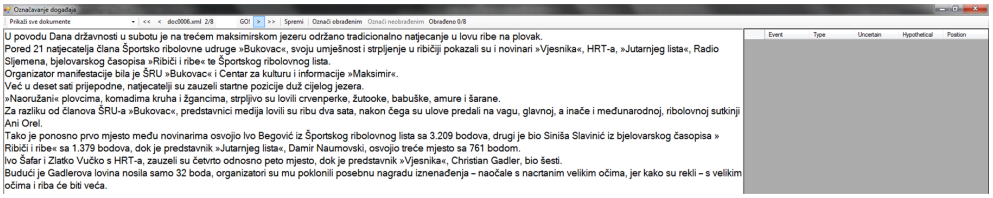
Ukoliko je događaj hipotetski, potrebno ga je označiti kao takvog. Dobar indikator hipotetskih događaja je korištenje modalnih glagola. Također, budući događaji označavaju se kao hipotetski jer se još nisu dogodili. Hipotetski karakter događaja određuje se samo iz već pročitanog dijela teksta. Drugim riječima, ukoliko se u nekoj rečenici sazna da se neki događaj u prethodnoj rečenici (a koji je tamo bio hipotetski) doista dogodio, taj prethodni događaj ostaje hipotetski u tom kontekstu.

*SDP se, nadalje, **zalaže** i za **dogovor** stranaka o cenzusu od 2,5 do 6 posto.*

***Rekao** je da će **otići** na more. Nakon što se **vratio** s mora, ništa nije bilo isto.*

Semantički razredi svrstani su u tri skupine prema prioritetu, od najvišeg do najnižeg:

1. REPORTING, ASPECTUAL, PERCEPTION
2. I_ACTION



Slika A.1: Početni izgled aplikacije za označavanje.

3. OCCURRENCE , HALF_GENERIC, STATE_CHANGE

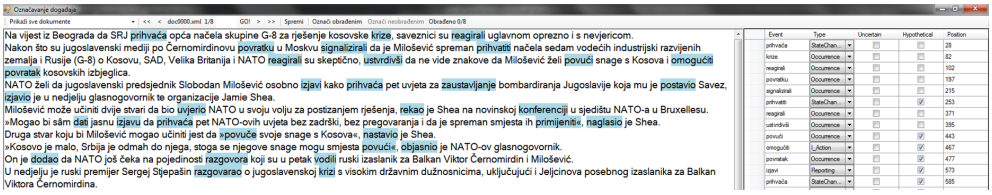
A.6. Aplikacija za označavanje događaja

Događaje je, prema prethodnim uputama, potrebno označavati pomoću aplikacije za označavanje događaja[1]. Glavna forma za označavanje događaja prikazana je na slici A.1. Na desnoj strani nalazi se popis označenih događaja u tekstu dokumenta koji je prikazan na lijevoj strani. Na početku obrade svakog pojedinog dokumenta popis događaja bit će prazan.

Označavanje riječi događajem obavlja se dvostrukim klikom na riječ. Potrebno je dva puta kliknuti bilo gdje unutar riječi koja se želi označiti kao događaj. Kada se riječ označi kao događaj, u tablici na desnoj strani ekrana pojavit će se zapis za taj događaj. U tom zapisu potrebno je označiti razred (OCCURRENCE, I_ACTION, REPORTING, PERCEPTION, ASPECTUAL, HALF_GENERIC, STATE_CHANGE) u koji označeni događaj spada i hipotetski karakter događaja (oznaka HYPOTHETICAL). Riječ odabrana kao nositelj događaja bit će označena svijetlo plavom pozadinom teksta (osim trenutno označenog događaja u listi događaja čija će pozadina biti svijetlo roza). Pojedino označavanje moguće je ukloniti ponovnim dvostrukim klikom na označenu riječ. Oznaka UNCERTAIN može se koristiti za rubne slučajeve ukoliko označivač nije uspio odrediti s nekom dozom sigurnosti određenu značajku događaja (je li riječ događaj, kojem razredu pripada, je li hipotetski ili realan). Oznake je moguće određivati pritiskom na tipku *Control* i prvo slovo oznake (npr. Ctrl+S za STATE_CHANGE). Prikaz izgleda aplikacije nakon označavanja događaja dan je na slici A.2.

Trenutno stanje označavanja dokumenta moguće je u svakom trenutku trajno pohraniti klikom na gumb "Spremi". Jednom kada je obrađen cijeli dokument (označeni su svi događaji u dokumentu), dokument je moguće označiti obrađenim (gumb "Označi

[1]Aplikacija za označavanje događaja izrađena je u sklopu doktorske disertacije mag. ing. comp. Gorana Glavaša koja se bavi ekstrakcijom događaja i vremenskih relacija iz tekstova na engleskom jeziku



Slika A.2: Aplikacija s otvorenim označenim tekstom.

obrađenim"). Označavanje dokumenta obrađenim služi prvenstveno označivaču kako bi vidio svoj napredak u označavanju (krajnje desno u alatnoj traci nalazi se broj označenih dokumenata i ukupan broj dokumenata za označavanje). Dok je dokument označen obrađenim nije ga moguće dalje obrađivati. Ukoliko, međutim, označivač uoči da je nešto propustio ili krivo označio, obrađeni dokument je moguće ponovno uređivati tako da ga se označi neobrađenim (klik na gumb "Označi neobrađenim"). Navigacija po dokumentima omogućena je gumbima "Prvi (|<)", "Prethodni (<)", "Sljedeći (>)", "Posljednji (>|)". Također je moguće skočiti na proizvoljni dokument u redoslijedu upisom rednog broja toga dokumenta i klikom na gumb "GO!".

A.7. Dodatni primjeri

U nastavku su dani još neki primjeri označavanja i objašnjenja pojedinih oznaka.
Na vijest iz Beograda da SRJ ***prihvaća*** *opća načela skupine G-8 za rješenje kosovske* ***krize****, saveznici su* reagirali *uglavnom oprezno i s nevjericom.“*

- događaj **prihvaća** označen je oznakom STATE_CHANGE jer se promijenilo “stanje prihvaćanja”;
- **kriza** (OCCURRENCE) označava složeni događaj koji u trenutku pisanja teksta još uvijek traje;
- **reagirali** (OCCURRENCE).

Nakon što su jugoslavenski mediji po Černomirdinovu ***povratku*** *u Moskvu* ***signalizirali*** *da je Milošević spreman prihvatiti načela sedam vodećih industrijski razvijenih zemalja i Rusije (G-8) o Kosovu, SAD, Velika Britanija i NATO* ***reagirali*** *su skeptično,* ***ustvrdivši*** *da ne* ***vide*** *znakove da Milošević želi* ***povući*** *snage s Kosova i* ***omogućiti*** ***povratak*** *kosovskih izbjeglica.*

- **povratku** (OCCURRENCE);

- **signalizirali** (OCCURRENCE) nije I_ACTION jer odmah iza njega ne slijedi događaj (*spreman* je stanje);
- **reagirali** (OCCURRENCE), **ustvrdivši** (OCCURRENCE);
- **vide** u ovom kontekstu predstavlja stanje;
- **povući** (OCCURRENCE), hipotetski;
- **omogućiti** (I_ACTION) se odnosi na **povratak** (OCCURRENCE), oba su hipotetska.

*NATO želi da jugoslavenski predsjednik Slobodan Milošević osobno **izjavi** kako **prihvaća** pet uvjeta za **zaustavljanje** **bombardiranja** Jugoslavije koja mu je **postavio** Savez, **izjavio** je u nedjelju glasnogovornik te organizacije Jamie Shea.“*

- **izjavi** (REPORTING), hipotetski;
- **prihvaća** (STATE_CHANGE), hipotetski;
- **zaustavljanje** (ASPECTUAL), hipotetski;
- **bombardiranja** (OCCURRENCE);
- **postavio** (OCCURRENCE);
- **izjavio** (REPORTING).

*Na **razgovorima** o narednim potezima Rusije glede **okončanja** krize **sudjelovali** su i ministar obrane Igor Sergejev, ministar vanjskih poslova Igor Ivanov i voditelj obavještajne službe za inozemstvo Vjačeslav Trubnikov, **rekao** je glasnogovornik vlade.“*

- **razgovorima** je HALF_GENERIC jer označava više različitih događaja, ali zna se na koje se razgovore misli;
- **okončanja** (ASPECTUAL), hipotetski;
- **krize** (OCCURRENCE);
- **sudjelovali** (OCCURRENCE);
- **rekao** (REPORTING).

Dodatak B
Upute za označavanje vremenskih relacija

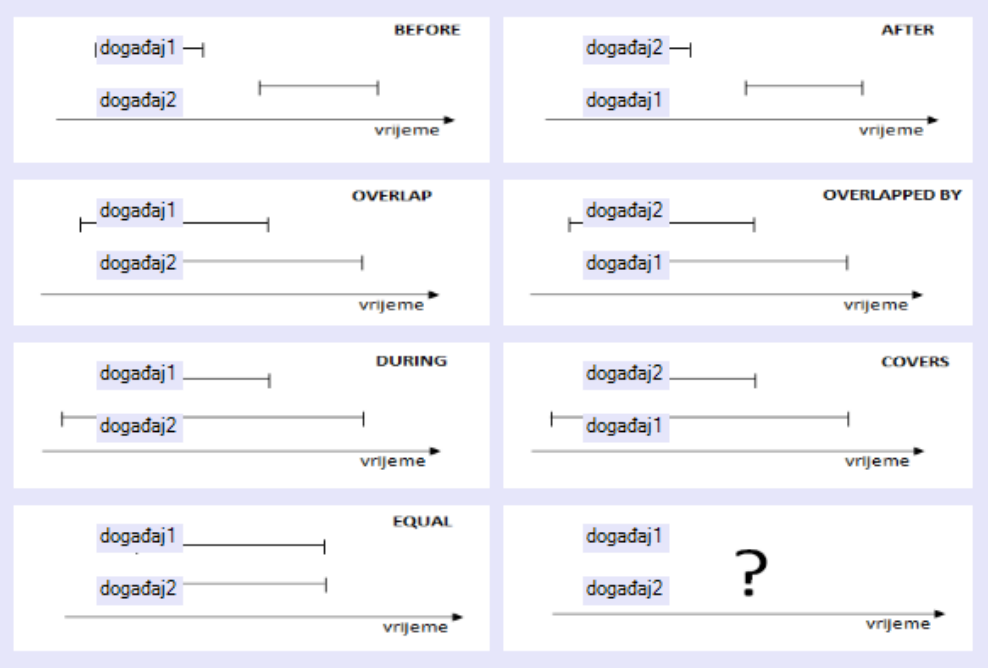
U drugoj fazi označavanja potrebno je označiti vremenske relacije između događaja u novinskim tekstovima na hrvatskom jeziku. U nastavku dokumenta dan je opis zadatka označavanja relacija. Taj opis obuhvaća objašnjenje pojma relacija u kontekstu zadatka, smjernice za označavanje te primjere označavanja pojedinih relacija.

B.1. Što je vremenska relacija?

Vremenska relacija opisuje odnos dva događaja u vremenu. U okviru ovog istraživanja razmatraju se vremenske relacije svih parova događaja koji se spominju unutar iste rečenice. Događaji vezani relacijom pritom predstavljaju uređeni par, tj. redoslijedom pojavljivanja u tekstu određeno je koji događaj je prvi član para, a koji drugi. Ukoliko se ispituje relacija para (*događaj1*, *događaj2*), tada se ne ispituje i relacija obrnutog para (*događaj2*, *događaj1*).

Pri označavanju je potrebno odrediti tip vremenske relacije. Tip relacije određen je odnosom početnih i završnih točaka događaja. Tipovi relacija prikazani su na slici B.1. U nastavku slijede pojašnjenja pojedinih tipova popraćena primjerima. Svaki primjer prikazuje rečenicu ili dio rečenice u kojoj se promatra samo jedna relacija, a događaji vezani promatranom relacijom **podebljani su i podcrtani**. Uz neke primjere dana su i dodatna pojašnjenja.

1. BEFORE — ovaj tip relacije obuhvaća sve relacije u kojima je prvi događaj u potpunosti prethodio drugom, tj. započeo je i završio prije početka drugog događaja. Iznimno, ovaj tip obuhvaća i one parove događaja gdje se trenutak završetka prvog događaja i trenutak početka drugog događaja podudaraju.



Slika B.1: Tipovi vremenskih relacija.

*Nakon što su jugoslavenski mediji po Černomirdinovu **povratku** u Moskvu **signalizirali** da je...*

*NATO želi da jugoslavenski predsjednik Slobodan Milošević osobno **izjavi** kako **prihvaća**...*

- uz pretpostavku da **izjavljivanje** označava službeni početak **prihvaćanja**, odgovarajući je tip BEFORE.

2. AFTER — ovim tipom relacije obuhvaćene su one relacije u kojima prvi događaj slijedi nakon drugog događaja, tj. započinje nakon završetka drugog događaja. Također, ovaj tip obuhvaća i one parove događaja gdje se trenutak završetka drugog događaja i trenutak početka prvog događaja podudaraju. Ovaj tip relacije jednak je tipu BEFORE uz zamjenu uloga prvog i drugog događaja.

   *...SAD, Velika Britanija i NATO **reagirali** su skeptično, **ustvrdivši** da ne vide znakove...*

   *NATO želi da jugoslavenski predsjednik Slobodan Milošević osobno **izjavi** kako prihvaća pet uvjeta za zaustavljanje bombardiranja Jugoslavije koja mu je **postavio***

Savez...

3. OVERLAP -– u ovaj tip ubrajaju se relacije u kojima je prvi događaj počeo prije početka drugog događaja, ali završio je nakon početka drugog događaja. Kod ovakvih relacija postoji određeni vremenski period u kojem su oba događaja trajala, tj. njihovo se trajanje djelomično preklapa.

   *Za vrijeme **racije** počeo je **bježati** i nije se zaustavio sve dok nakon tri dana nije izišao iz zemlje.*

4. OVERLAPPED BY – u ovaj tip relacije ubrajaju se relacije u kojima je drugi događaj počeo prije početka prvog događaja, ali završio je nakon početka prvog događaja. Kod ovakvih relacija postoji određeni vremenski period u kojem su oba događaja trajala, tj. njihovo trajanje se djelomično preklapa. Ovaj tip relacije odgovara tipu OVERLAP uz prethodnu zamjenu uloga prvog i drugog događaja.

   *Lokalne paravojne skupine nastavile su **zlostavljati** stanovništvo dugo nakon završetka **rata**.*

5. DURING -– ovom tipu relacije pripadaju parovi događaja kod kojih je trajanje prvog događaja u potpunosti obuhvaćeno trajanjem drugog događaja. Drugim riječima, prvi događaj započeo je u trenutku započinjanja ili nakon početka drugog događaja te je završio prije završetka ili u trenutku završetka drugog događaja.

   *NATO želi da jugoslavenski predsjednik Slobodan Milošević osobno **izjavi** kako prihvaća pet uvjeta za zaustavljanje **bombardiranja**...*

   *NATO želi da jugoslavenski predsjednik Slobodan Milošević osobno izjavi kako prihvaća pet uvjeta za **zaustavljanje** **bombardiranja**...*

6. COVERS – ovom tipu relacije pripadaju parovi događaja kod kojih je trajanje drugog događaja u potpunosti obuhvaćeno trajanjem prvog događaja. Drugim riječima, drugi događaj započeo je u trenutku započinjanja ili nakon početka prvog događaja te je završio prije završetka ili u trenutku završetka prvog događaja. Ovaj tip relacije jednak je tipu DURING uz prethodnu zamjenu uloga prvog i drugog događaja.

*Na vijest iz Beograda da SRJ prihvaća opća načela skupine G-8 za rješenje kosovske **krize**, saveznici su **reagirali** uglavnom oprezno i s nevjericom.*

*...pet uvjeta za zaustavljanje **bombardiranja** Jugoslavije koja mu je **postavio** Savez...*

- uvjeti za zaustavljanje **postavljeni** su nakon što je **bombardiranje** već počelo.

7. EQUAL — ovaj tip relacije odnosi se na sve relacije čiji su događaji počeli i završili u jednakim trenutcima. Ovo je poseban slučaj tipova DURING i COVERS. Najčešće se ovom relacijom povezuju riječi koje se odnose na isti događaj.

   *On je dodao da NATO još čeka na pojedinosti **razgovora** koji su u petak **vodili** ruski izaslanik za Balkan Viktor Černomirdin i Milošević.*

   *Dodao je kako se **sastanak** **održao** po Jeljcinovoj naredbi.*

8. UNKNOWN -– posljednji tip relacije pridjeljuje se svim parovima događaja čiji je odnos nemoguće odrediti, primjerice u slučaju dva moguća hipotetska događaja za koje nisu određena točna vremena kad bi se mogli ostvariti ili općenito za bilo koja dva događaja čija vremena nisu definirana dovoljno dobro za određivanje relacije.

   *Roko Karanušić je **eliminiran** je sa 6-7 (5), 6-2, 6-3, a kod tenisačica, u 1. kolu **ispale** su i Ivana Abramović i Dora Krstulović.*

   - - iz teksta nije poznato koji meč je bio prije.

B.2. Kako označavati?

U ovom dijelu dane su neke dodatne smjernice za označavanje relacija.

1. Ukoliko je barem jedan događaj hipotetski, potrebno je u postupku određivanja vremenske relacije primijeniti tzv. analizu mogućih svjetova (engl. *possible worlds analysis*), prema (Bethard et al., 2007). Drugim riječima, svaki hipotetski događaj potrebno je smjestiti u odgovarajuće vremenske okvire kao da se on

doista i ostvario. Iznimka je slučaj kada su događaji kontekstom definirani kao strogo isključivi, pri čemu im se dodjeljuje tip relacije UNKNOWN.

*SAD, Velika Britanija i NATO **reagirali** su skeptično, ustvrdivši da ne vide znakove da Milošević želi povući snage s Kosova i omogućiti **povratak** kosovskih izbjeglica.*

- **povratak** je hipotetski događaj koji bi se u slučaju ostvarivanja dogodio u budućnosti, nakon **reagiranja**, pa je BEFORE odgovarajući tip relacije.

*NATO želi da jugoslavenski predsjednik Slobodan Milošević osobno **izjavi** kako prihvaća pet uvjeta za zaustavljanje **bombardiranja**...*

- dva hipotetska događaja, tip relacije je BEFORE, jer je za zaustavljanje **bombardiranja** potrebno prihvatiti uvjete, a uvjeti se službeno prihvaćaju **izjavom** o prihvaćanju.

2. U skupu relacija česti su parovi događaja od kojih jedan pripada razredu I_ACTION, a obje riječi odnose se na isti događaj. Pojavljivanje takvog para obično podrazumijeva tip relacije EQUAL.

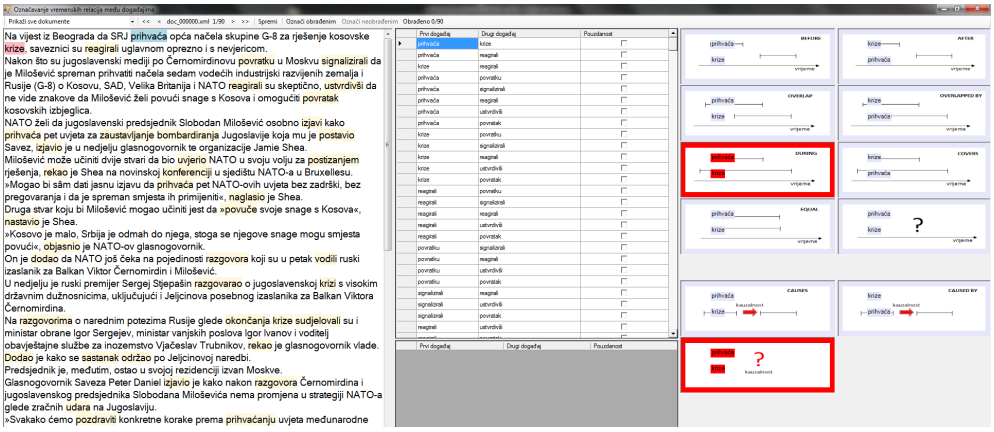
   *On je dodao da NATO još čeka na pojedinosti **razgovora** koji su u petak **vodili** ruski izaslanik za Balkan Viktor Černomirdin i Milošević.*

   *Dodao je kako se **sastanak** **održao** po Jeljcinovoj naredbi.*

3. U slučaju događaja poput dozvoliti, omogućiti, zaboraviti itd. koji podrazumijevaju promjenu stanja i daljnje trajanje tog stanja (pri čemu to trajanje može biti beskonačno) potrebno je pri dodjeli tipa relacije uzeti u obzir samo trenutak te promjene, tj. početni trenutak tog događaja.

   *Na vijest iz Beograda da SRJ **prihvaća** opća načela skupine G-8 za rješenje kosovske **krize**, saveznici su reagirali uglavnom oprezno i s nevjericom.*

   (a) tip relacije je DURING jer se uzima u obzir samo početak prvog događaja (**prihvaća**), a u trenutku prihvaćanja drugi događaj (**kriza**) je već trajao.



Slika B.2: Početni izgled aplikacije za označavanje vremenskih relacija.

*Na vijest iz Beograda da SRJ **prihvaća** opća načela skupine G-8 za rješenje kosovske krize, saveznici su **reagirali** uglavnom oprezno i s nevjericom.*

(a) - tip relacije je BEFORE jer je do **reagiranja** došlo tek nakon **prihvaćanja**, pri čemu se opet uzima u obzir samo početak prvog događaja.

B.3. Aplikacija za označavanje vremenskih relacija

Relacije je, prema prethodnim uputama, potrebno označavati pomoću aplikacije za označavanje relacija[1]. Glavna forma za označavanje relacija prikazana je na slici B.2. Forma se sastoji od tri dijela. Lijevi dio forme sadrži tekst dokumenta s događajima označenim žutom bojom. U središnjem dijelu forme prikazane su sve vremenske relacije. U desnom dijelu forme nalaze se kontrole za označavanje relacija.

Označavanje relacije započinje odabirom željene relacije u središnjem dijelu forme. Kad se odabere relacija, aplikacija u lijevom dijelu posebno ističe događaje relacije tako da prvi događaj označi plavom, a drugi crvenom bojom. Također, u desnom dijelu forme u zasebnim poljima prikazuju se tipovi vremenskih relacija (BEFORE, AFTER, OVERLAP, OVERLAPPED BY, DURING, COVERS, EQUAL i UNKNOWN). Unutar svakog polja dan je jednostavan grafički prikaz tipa vremenske relacije. Također se prikazuju i oznake kauzalnosti, no one se ovom istraživanju ne razmatraju i potrebno ih je zanemariti. Trenutno odabrana relacija istaknuta je crvenom bojom. Oznaka relacije

[1] Aplikacija za označavanje relacija izrađena je u sklopu doktorske disertacije mag. ing. comp. Gorana Glavaša koja se bavi ekstrakcijom događaja i vremenskih relacija iz tekstova na engleskom jeziku

mijenja se klikom na odgovarajuće polje.

Trenutno stanje označavanja dokumenta moguće je u svakom trenutku trajno pohraniti klikom na gumb "Spremi". Jednom kada je obrađen cijeli dokument (označeni su svi događaji u dokumentu), dokument je moguće označiti obrađenim (gumb "Označi obrađenim"). Označavanje dokumenta obrađenim služi prvenstveno označivaču kako bi vidio svoj napredak u označavanju (krajnje desno u alatnoj traci nalazi se broj označenih dokumenata i ukupan broj dokumenata za označavanje). Dok je dokument označen obrađenim, nije ga moguće dalje obrađivati. Ukoliko, međutim, označivač uoči da je nešto propustio ili krivo označio, obrađeni dokument je moguće ponovno uređivati tako da ga se označi neobrađenim (klik na gumb "Označi neobrađenim"). Navigacija po dokumentima omogućena je gumbima "Prvi («)", "Prethodni (<)", "Sljedeći (>)", "Posljednji (»)". Također je moguće skočiti na proizvoljni dokument u redoslijedu upisom rednog broja toga dokumenta i klikom na gumb "GO!".

Ekstrakcija događaja i vremenskih relacija u tekstovima na hrvatskome jeziku

Sažetak

Danas su dostupne goleme količine pisanoga teksta koje predstavljaju velik izvor znanja. Automatska ekstrakcija informacija iz tekstnih podataka, poput ekstrakcije događaja i vremenskih relacija među događajima, omogućava iskorištavanje tog znanja u različitim područjima ljudske djelatnosti. Ekstrakcija događaja i vremenskih relacija netrivijalni su zadatci obrade prirodnog jezika i predmetom su intenzivnog istraživanja.

U okviru ovog istraživanja proučeni su postupci za ekstrakciju događaja i vremenskih relacija temeljeni na metodama strojnog učenja. Razrađen je postupak za ekstrakciju događaja i vremenskih relacija u tekstovima na hrvatskom jeziku. Provedeno je označavanje odgovarajućeg tekstnog uzorka i odabrane su najprikladnije značajke uzevši u obzir ograničenost jezičnotehnoloških alata za hrvatski jezik. Provedeno je eksperimentalno vrednovanje točnosti ekstrakcije uporabom različitih metoda strojnog učenja, analiza značajki i analiza pogrešaka. Dobiveni rezultati su obećavajući, uz postignutu F_1-mjeru od 93% pri označavanju događaja, 77% pri označavanju semantičkih razreda događaja te 64% pri označavanju vremenskih relacija.

Ključne riječi: ekstrakcija informacija, obrada prirodnog jezika, događaj, vremenska relacija, klasifikacija, strojno učenje, hrvatski jezik

Event and temporal relation extraction in Croatian language texts

Abstract

There are large amounts of written text available which present a great source of knowledge. Automatic information extraction from textual data, such as event extraction and temporal relation extraction, enables the use of such knowledge in different areas of human activity. Event extraction and temporal relation extraction are nontrivial natural language processing tasks and, as such, are the object of extensive research.

In this paper different approaches to event and temporal relation extraction were studied. A method was devised for event and temporal relation extraction in Croatian language texts. An adequate text sample annotation was performed and, given the limited availability of linguistic tools for Croatian, the most appropriate features were selected. Experimental evaluation was conducted which yielded promising results, producing the F-score of up to 93% for event extraction, 77% for event type classification and 64% for temporal relation extraction.

Keywords: information extraction, natural language processing, event, temporal relation, classification, machine learning, Croatian language