Automatska obrada vremenskih izraza srpskog jezika

Jelena Jaćimović

Filološki fakultet Univerziteta u Beogradu

Mentor

Prof. dr Cvetana Krstev

Seminar Društva za jezičke tehnologije i resurse
Matematički fakultet Univerziteta u Beogradu
24.12.2015. godine

Pojam vremena

- Vreme igra ulogu univerzalnog referentnog sistema koji se koristi za povezivanje,ređanje, merenje i upoređivanje intervala događaja i stanja
- Značenje vremena u svakodnevnom životu i način na koji čovek poima vreme ogledaju se i u komunikaciji, pre svega u jezičkim izrazima koji se učestalo koriste u svakodnevnom govoru

Obrada vremenskih informacija

- Napredni računarski alati za automatsku obradu tekstova, sposobni da automatski označe i informacije koje referišu na vreme

  – doprineti očuvanju srpskog jezika u digitalnom okruženju

  – uticati na poboljšanje učinka primene mnogih drugih aplikacija iz oblasti jezičkih tehnologija

Vreme u prirodnom jeziku

- Iako ljudi u komunikaciji nemaju nikakav problem u razumevanju vremena i vremenskih odnosa, njihovo formalizovanje na jezik razumljiv računarima predstavlja težak zadatak
  - ista vremenska informacija može biti saopštena na različite načine
  - vremenska informacija ne mora uvek da bude izrečena eksplicitno, već često može da bude implicitna i da zahteva tumačenja ili zaključke izvedene na osnovu znanja koje posedujemo o svetu

Vremenske informacije

- Srpski jezik, kao i bilo koji drugi prirodni jezik, poseduje nekoliko mehanizama za izražavanje vremenskih informacija, koje uopšteno mogu biti grupisane u tri velike kategorije:
  - vremenski izrazi,
  - događaji i
  - vremenski odnosi koji postoje između njih.

Resursi i računarski pristupi

- Različiti pristupi projektovanju sistema za ekstrakciju vremenskih izraza (MUC, Mani i Vilson, TERN, TempEval)
- Anotirani korpusi (TERN, TimeBank, AQUAINT, TempEval korpusi)
- Sheme za obeležavanje vremenskih izraza (TIMEX, TIMEX2, STAG, TIMEX3)
  - ISO 24617-1:2012 Language resource management -- Semantic annotation framework (SemAF) -- Part 1: Time and events (SemAF-Time, ISO-TimeML)

Vremenski izrazi

- Vremenski izrazi, kao fraze prirodnog jezika koje se direktno odnose na vreme, pružaju informaciju o tome kada se nešto dogodilo, koliko dugo je trajalo ili koliko često se dešava

15. aprila 1999. godine, juče, 8 h, sinoć

osam sati, nekoliko dana

mesečno, svake srede, dva puta godišnje

Cilj

- formalizacija vremenskih izraza i njihovo automatsko obeležavanje u nestrukturiranim tekstovima srpskog jezika sa postizanjem visokog nivoa odziva i preciznosti
  - utvrđivanje tipologije semantičkih obeležja i semantičkih tipova vremenskih izraza koji se pojavljuju u tekstovima srpskog jezika
  - prepoznavanje vremenskih izraza
  - normalizacija njihovih vrednosti

Prepoznavanje vremenskih izraza

Korpus

- Kao tekstualni resurs korišćen je deo Korpusa savremenog srpkog jezika, odnosno kolekcija novinskih tekstova prikupljenih tokom 2005-2012. godine iz više različitih izvora na srpskom jeziku (Glas javnosti, Blic, Večernje novosti, Srpski nacional, Politika, Danas, B92, Beta, Tanjug, FoNet)

Identifikacija vremenskih izraza

- otkrivanje jezičkih struktura reprezentovanih različitim formalnim jedinicama, kojima se prenose tri osnovna tipa vremenskog značenja:
  - KAD (pravo vreme),
  - KOLIKO DUGO (kvantitativnost u vremenskom smislu) i
  - KOLIKO ČESTO (iterativnost).

Leksički okidači

Vrsta reči	Leksički okidač
Imenice	sekund, minut, sat, dan, vikend, nedelja, mesec, godina, decenija, vek, podne, noć, ponedeljak, mart, proleće
Vlastite imenice	Božić, Nova godina, Uskrs
Posebni vremenski obrasci	12:35, 3.04.1999., 11/30/2005, 1998, 1970-ih
Pridevi	prošli, tekući, sledeći, mesečni, devedeseti
Prilozi	mesečno, dnevno, nedeljno, večeras, danas, juče, noćas, jesenas, zimi, sada, tada, onda, sedamdesetih
Broj	4, dva, prvog, 5.

Određivanje tipa

- Vremenski izrazi koji impliciraju tačku u vremenu
  - kalendarski datumi (DATE)
  - vremena dana (TIME)
- Vremenski izrazi koji impliciraju trajanje (DURATION)
- Vremenski izrazi koji impliciraju učestalost ponavljanja vremena (SET)
  - povremena ponavljanja
  - regularna ponavljanja

Granularnost vremenskih izraza

- Mogu biti predstavljeni različitim nivoima detaljnosti

2. milenijum

XIII veka

1970-tih godina

2011. godine

februaru 2009.

17. nedelju 2009.

18. januara 2014. godine

15 časova 21. januara 2003. godine

7:30 1. jula 2004. godine

05:37:39, 19.01.2010.

Vremenski izrazi

- izrazi čije vrednosti mogu biti normalizovane na osnovu njih samih (**potpuno precizna**, **kontekstno nezavisna** ili **apsolutna vremena**)

  2007. godine, 4. aprila 1999.

- izrazi koji zahtevaju vrednost drugog vremenskog izraza kao orijentira u procesu normalizacije (**nedovoljno precizna**, **kontekstno zavisna** ili **relativna vremena**)

  prošle godine, juče

Određivanje opsega

- pun opseg izraza obuhvaćenog etiketom obeležavanja mora da bude jedna od gramatičkih kategorija:
  - imenica (npr. *danas, petak*)
  - imenička sintagma (npr. *sredu uveče, prošle godine*)
  - pridev (npr. *današnji*)
  - prilog (npr. *letos, mesečno*)
  - pridevska/priloška sintagma (npr. *taj godišnji, rano jutros*)

Određivanje opsega

- Za razliku od predloga i veznika, reči ili grupe reči koje na određeni način modifikuju ili kvantifikuju vremenski izraz biće uključene u pun opseg koji je potrebno obeležiti.

početkom *godine,* ***sredinom*** *1999.*

manje od *dva sata,* ***gotovo*** *dva veka*

Format za obeležavanje

- umetanjem <TIMEX3> etikete u okviru koje će biti definisani atributi:

```
type::='DATE'|'TIME'|'DURATION'|'SET'

temporalFunction::='true'|'false'
```

Atribut `type`

- specifikuje semantičku klasu prepoznatog izraza

```
<TIMEX3 type="DATE">juče</TIMEX3>
<TIMEX3 type="TIME">18 časova</TIMEX3>
<TIMEX3 type="DURATION">dva sata</TIMEX3>
<TIMEX3 type="SET">svakog petka</TIMEX3>
```

Atribut `temporalFunction`

- binarni atribut koji ukazuje na to da li se radi o apsolutnom ili relativnom vremenskom izrazu

```
<TIMEX3 type="DATE"
temporalFunction="false">27.
marta 2005.</TIMEX3>
<TIMEX3 type="DATE"
temporalFunction="true">prošle
godine</TIMEX3>
```

Metod za prepoznavanje

- Izvori za automatsku obradu srpskog jezika razvijeni su putem metode konačnih stanja
- Pomoću grafičkog korisničkog interfejsa Unitex-a i radnog okruženja koje ovaj sistem obezbeđuje izvodi se ceo proces, od prethode obrade teksta, preko kreiranja pravila prepoznavanja, do samog obeležavanja, odnosno izdvajanja vremenskih informacija

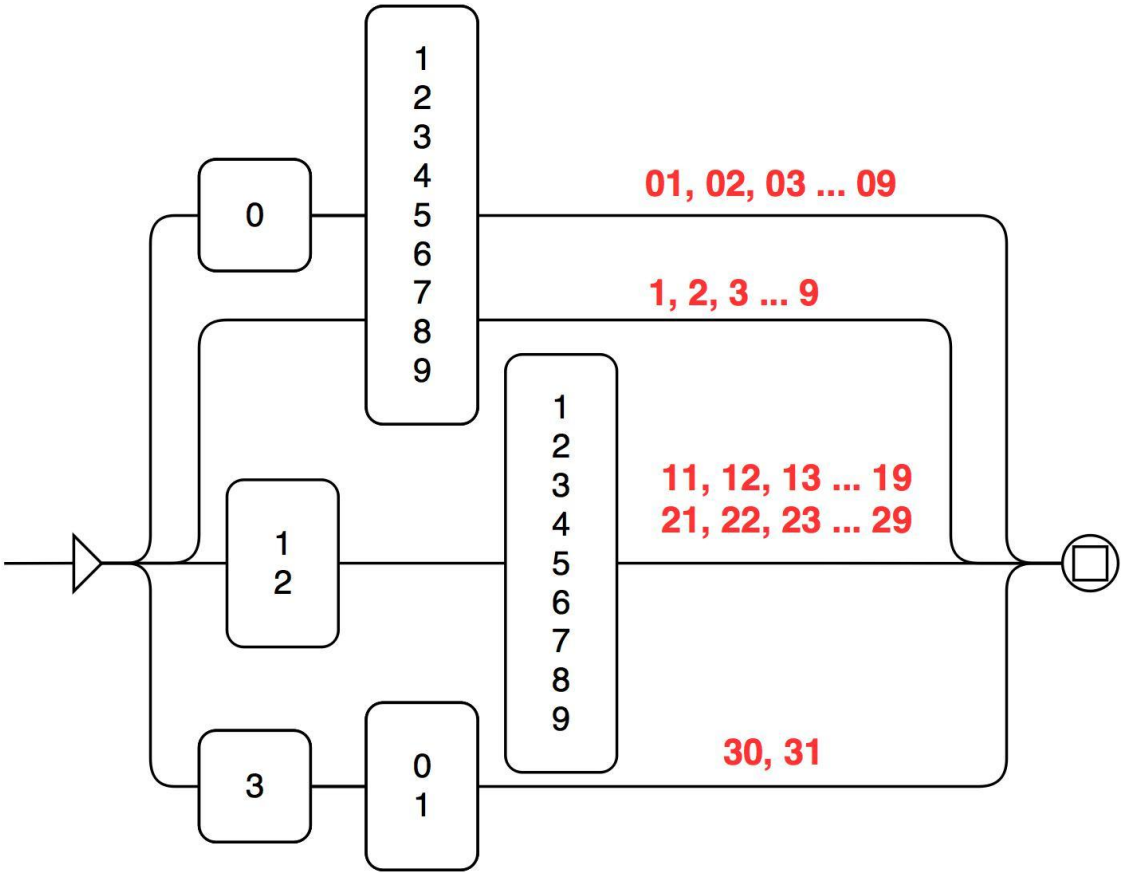
Metoda konačnih stanja

- Zgodni za zadatke plitkog parsiranja, moguće je postići dobre rezultate korišćenjem pravila u obliku konačnih automata
- Čitljivi i zgodni za korišćenje i lingvistima

Konačni automati

- definišu formalni jezik na osnovu koga se utvrđuje da li određena niska pripada jeziku opisanom tim automatom ili ne

Konačni automat

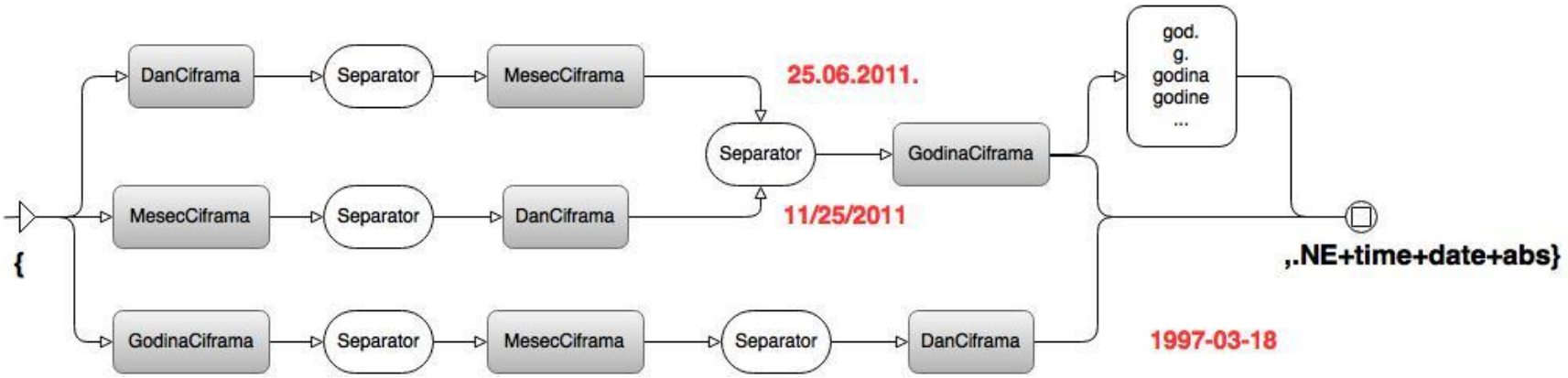


Graf *DanCiframa.grf* koji prepoznaje dane u mesecu iskazane arapskim brojevima

Konačni transducktori

- konačni transducktori definišu relacije između dva skupa niski karaktera, odnosno transformišu zadatu nisku u drugu nisku nad istom ili nekom drugom azbukom
- prepoznaju jednu nisku karaktera, a generišu neku drugu

Konačni transdukor



Transduktor *Cal_date.grf* za prepoznavanje i obeležavanje kalendarskih datuma iskazanih numeričkim obrascima

{13. juna 2008. godine,.NE+time+date+abs}

Kaskadna primena transduktora

- uzastopnu primenu serije transduktora na tekst preciznim redosledom kako bi se transformisao tekst ili ekstrahovali obrasci iz teksta
- Sistem za prepoznavanje vremenskih izraza srpskog jezika zasniva se na kaskadi transduktora – CasSys, koja je integrisana u Unitex sistem

Kaskada transduktora

- kaskada za prepoznavanje se sastoji od 16 transdukcija
- uloga identifikacija izraza, kao i određivanje opsega i tipa svakog otkrivenog izraza, a u skladu sa TimeML shemom (DATE, TIME, DURATION i SET)

Određeni redosled

- prioritet je dat transduktorima koji pronalaze najduže obrasce kako bi se izbegli slučajevi pogrešnog obeležavanja

2014. godine

15. mart 2014. godine

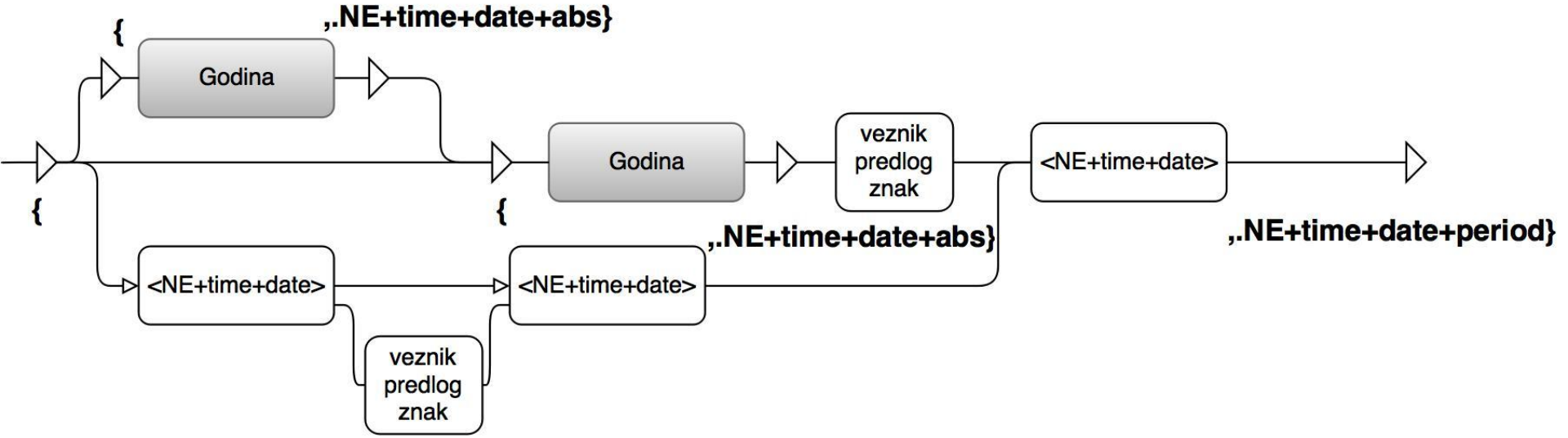
mart 2014. godine

- dvosmislene situacije

  Dao je tri gola u *dva sata* igre.

  Sastanak počinje u *dva sata*.

Leksička etiketa



1989, 1999. *i 2000. godine*
7, 8. *i 9. aprila prošle godine*
2007*/2008. godine*
od ***11.*** *do 22. marta*
11. *ili 12. februara 2005.*
između ***petka 7.*** *i subote 8. novembra 2011.*

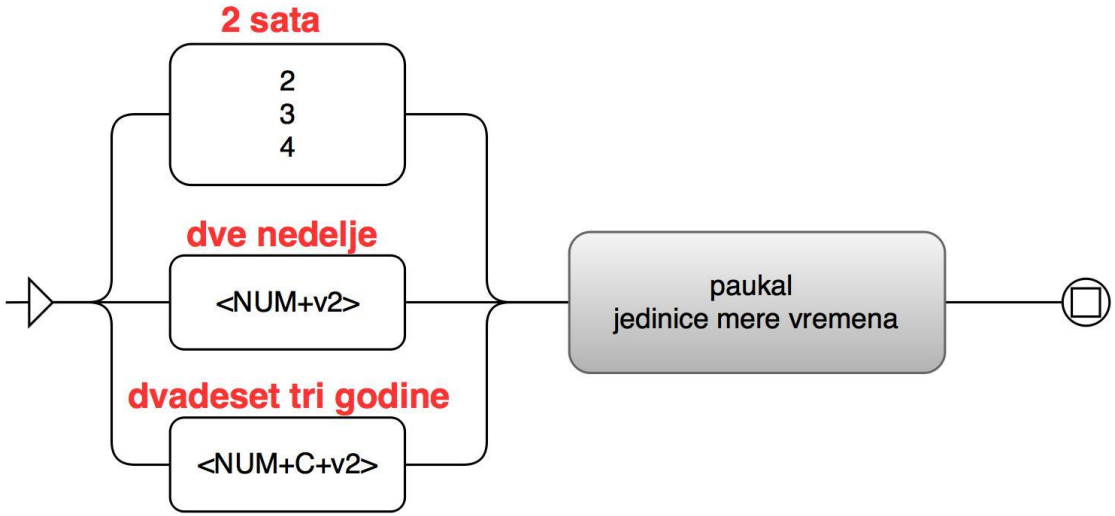
Leksička etiketa

- otkrivanje mnogo složenijih izraza, poput onih koji označavaju vremenski period ili predstavljaju kombinaciju kalendarskih datuma i vremena kao delova dana

od {{8. marta,.NE+time+date+rel} do {7. aprila,.NE+time+date+rel},.NE+time+date+period}

{{15. marta,.NE+time+date+rel} {oko 2 sata,.NE+time+hour+abs},.NE+time+hour}

Upotreba rečnika



Neke od putanja grafa zaduženog za prepoznavanje izraza koji ukazuju na trajanje

1.702 godine

1.702,.NUM+C+v2+NVAL=1702

<NUM+C+v2>

Izlaz transduktora

- na kraju procesa prepoznavanja vremenskih izraza sve leksičke etikete će biti konvertovane u XML etikete, odnosno etikete koje su u skladu sa TimeML shemom

{hiljadu i 200 godina,.NE+time+duration+abs}

```
<time.duration.abs>hiljadu i 200 godina</time.duration.abs>
<TIMEX3 type="DURATION" temporalFunction="false">hiljadu i 200 godina</TIMEX3>
```

Normalizacija vremenskih izraza

Normalizacija vremenskih izraza

- zadatak interpretacije vremenskih izraza u standardizovanom obliku (ISO 8601)

- usmerena je na utvrđivanje apsolutne vrednosti vremenskog izraza, bez obzira na jezički oblik kojim je iskazan

Format za obeležavanje

- više atributa definisanih TimeML uputstvom, koji se direktno odnose na konačnu interpretaciju vrednosti identifikovanih vremenskih izraza, biće uključeno u <TIMEX3> etiketu:
  - `value`
  - `mod`
  - `valueFromFunction`
  - `quant`
  - `freq`

Atribut `value` – DATE

- obavezni atribut, kojim se reprezentuju konačne vrednosti vremenskih izraza u standardizovanom obliku

YYYY-MM-DD

value=“2011-12-25“

proleća 2009. → value=“2009-SP“

22. novembra → value=“XXXX-11-22“

22. novembra prošle godine → value=“XXXX-11-22“

petak → value=“XXXX-WXX-05“

juče → value=“XXXX-XX-XX“

Atribut `value – TIME`

Thh:mm

osam i trideset uveče → value=“T20:30“

jutros → value=“TMO“

noćas → value=“TNI“

YYYY-MM-DDThh:mm

15. časova 30. decembra 2014. godine → value=“2014-12-30T15:00“

Atribut `value` – DURATION

Jedinica mere vremena	Kod za reprezentaciju
era	BC
milenijum	L
vek, stoleće	C
decenija	E
godina	Y
mesec	M
nedelja, sedmica	W
vikend	WE
dan	D
sat	H
minut	MIN
sekund	S

Pn(kod)

20 minuta → value="P20MIN"

9 meseci → value="P9M"

celu deceniju → value="P1E"

dva mlenijuma → value="P2L"

dve noći → value="P2NI"

poslednjih nekoliko godina → value="PXY"

sledećih par dana → value="PXD"

mesecima → value="PXM"

nakon godina → value="PXY"

Atribut `value` – SET

- Da bi vremenski izrazi koji ukazuju na učestalost ponavljanja bili u potpunosti anotirani, osim atributa `value`, neophodna je i upotreba atributa `quant` ili `freq`

dva puta nedeljno → PnW → value="P1W"

svakog oktobra → YYYY-MM → value="XXXX-10"

Atribut mod

- opcioni atribut
- prikazivanje značenja vremenskih izraza koji su na neki način kvantifikovani (npr. ***približno*** *2 sata,* ***ne više od*** *10 minuta*) ili modifikovani (npr. ***početkom*** *2005. godine,* ***krajem*** *aprila meseca*)

```
<TIMEX3 type="DURATION"
temporalFunction="false" value="P2H"
mod="APPROX">oko dva sata</TIMEX3>
```

Atribut mod

Tip vremenskog izraza	Token	Primeri izraza
Trajanje	LESS_THAN	*manje od* dva sata, *skoro* 5 minuta
	MORE_THAN	*više od* 2 sata
Tačka u vremenu i trajanje	START	*početkom* 2005, *prva polovina* godine
	MID	*sredinom* meseca
	END	*krajem* aprila 2006.
	APPROX	*oko* dva sata, *oko* 5. maja

Atribut `valueFromFunction`

- značajan je za kasniji proces računanja apsolutne vrednosti izraza pomoću drugih izraza koji će poslužiti kao orijentir
- sastoji se od:
  - oznake za računsku operaciju koju je potrebno primeniti (plus, minus ili znak jednakosti),
  - oznake za količinu (n≥0) jedinica vremenske mere koja treba da bude dodata ili oduzeta od vrednosti vremenskog izraza koji je uzet za orijentir
  - oznake za jedinicu mere vremena

```
<TIMEX3 type=”DATE” temporalFunction=”true” value=”XXXX-XX-XX” valueFromFunction=“-1D“>juče</TIMEX3>
```

Atribut `quant i freq`

- Atribut `quant` se prevashodno koristi za reprezentaciju vremenskih izraza koji ukazuju na regularnu učestalost ponavljanja perioda, a njegova vrednost je iskazana engleskim terminom izraza kojim se kvantifikuje vrednost

```
<TIMEX3 type="SET" value="P2D"
quant="EVERY">svaka 2 dana</TIMEX3>
<TIMEX3 type="SET" value="XXXX-10"
quant="EVERY">svakog
oktobra</TIMEX3>
```

Atribut `quant i freq`

- Atribut `freq` sadrži vrednost celog broja i nivo granularnosti frekvencije ponavljanja vremenskog izraza

```
<TIMEX3 type="SET" value="P1W"
freq="2X">dva puta
nedeljno</TIMEX3>

<TIMEX3 type="SET" value="P1W"
quant="EVERY" freq="3D">3 dana
svake nedelje</TIMEX3>
```

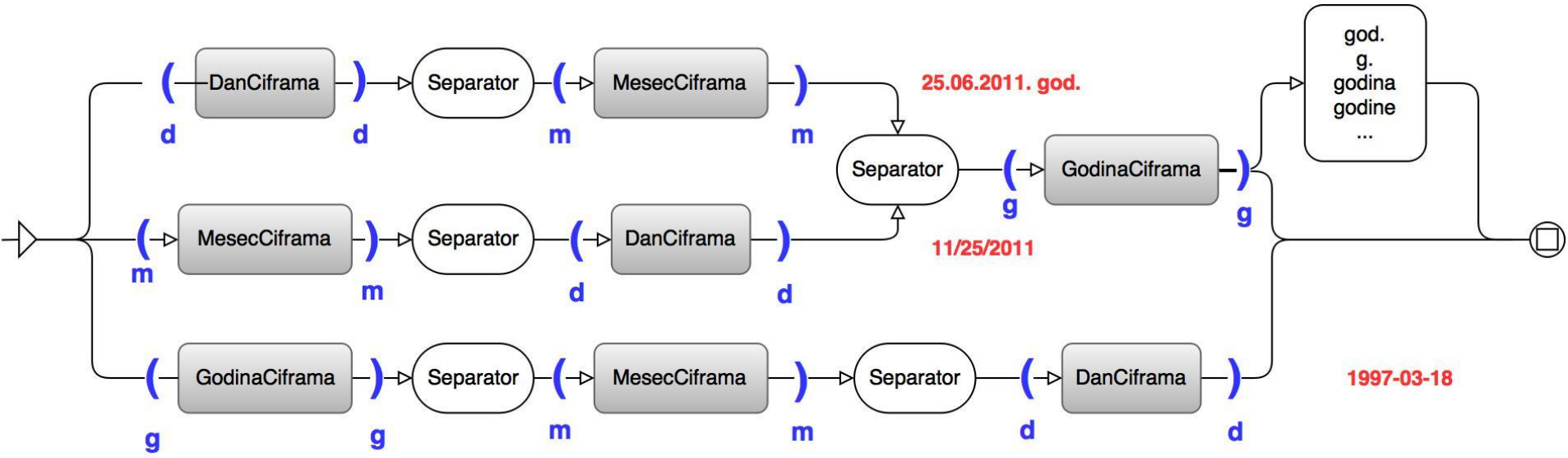
Lokalna gramatika za normalizaciju

- primenjuju se gotovo ista pravila korišćena za prepoznavanje vremenskih izraza
- sastoji se od četiri glavna transduktora, od kojih svaki odgovara postojećim semantičkim klasama vremenskih izraza jer od njih zavisi izgled normalizovane vrednosti

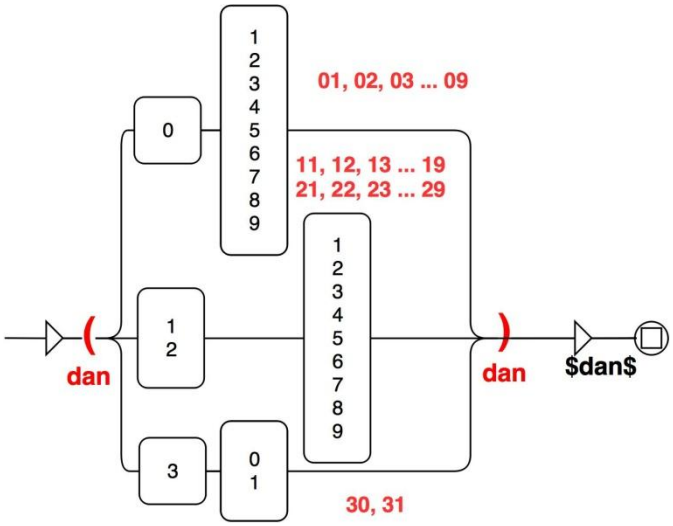
Lokalna gramatika za normalizaciju

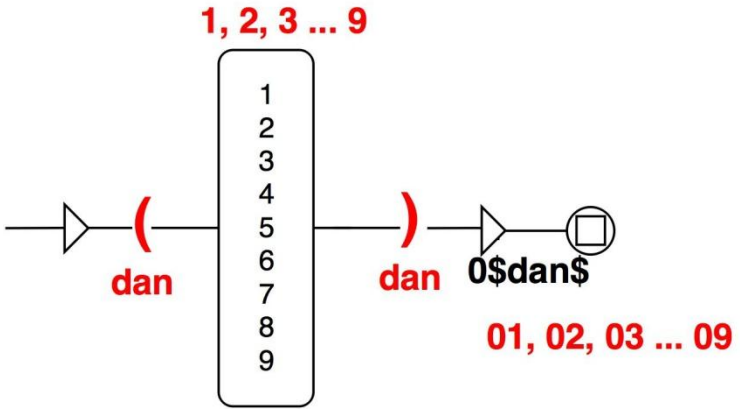
1. Datum (normalizuje vrednosti prepoznatih apsolutnih i relativnih vremenskih izraza koji impliciraju tačku u vremenu i u obliku kalendarskog datuma su)
2. Vreme dana (normalizuje vrednosti prepoznatih apsolutnih i relativnih vremenskih izraza koji impliciraju tačku u vremenu i u obliku vremena dana su)
3. Trajanje (normalizuje vrednosti prepoznatih apsolutnih i relativnih vremenskih izraza koji impliciraju trajanje)
4. Učestalost (normalizuje vrednosti prepoznatih vremenskih izraza koji impliciraju regularnu ili povremenu učestalost ponavljanja)

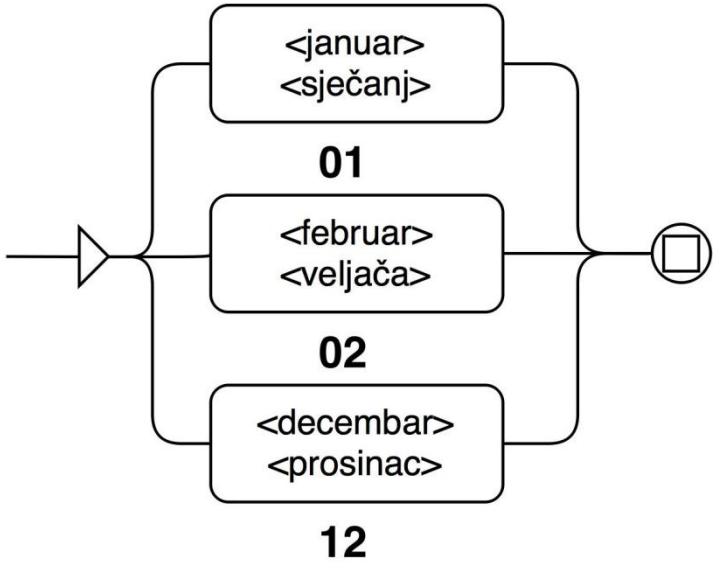
Lokalna gramatika za normalizaciju



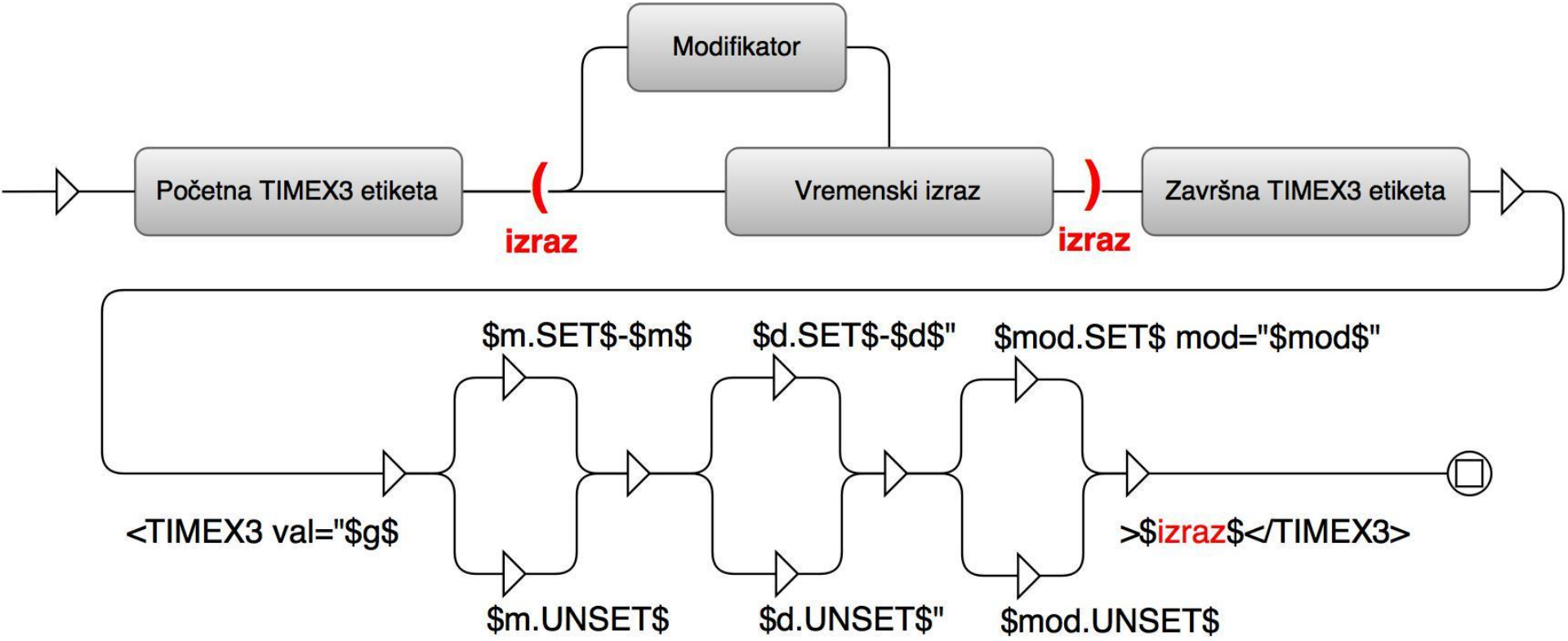
Lokalna gramatika za normalizaciju







Lokalna gramatika za normalizaciju



početkom 2009. godine

<TIMEX3 type="DATE" temporalFunction="false" value="2009" mod="START">početkom 2009. godine</TIMEX3>

Lokalna gramatika za normalizaciju

prošle godine

`value=„XXXX“`

`valueFromFunction=„-1Y“`

tokom *prošle godine*

`value=„P1Y“ valueFromFunction=„-1Y“`

godina,.JMV+XVAL=XXXX+KOL=Y

<JMV> $j.CODE.ATTR=XVAL$

<JMV> $j.CODE.ATTR=KOL$

Primena u domenu medicine

Vreme u medicini

- Vremenska dimenzija predstavlja osnovu pravilne interpretacije informacija koje se nalaze u medicinskim/kliničkim tekstovima
- Prevashodno su razvijani sistemi koji obrađuju strukturirane podatke
- Većina sistema je namenjena primeni u oblasti medicine

Cilj

- automatsko obeležavanje vremenskih izraza u nestrukturiranim medicinskim tekstovima srpskog jezika sa postizanjem visokog nivoa odziva i preciznosti
- procena efikasnosti metode za prepoznavanje vremenskih izraza u domenu medicinskih nestrukturiranih tekstova srpskog jezika

Deidentifikacija

- otkrivanje i uklanjanje, odnosno modifikovanje svih onih eksplicitno iskazanih ličnih podataka, koji se odnose na pacijenta (eng. *Protetcted Health Information*), a nalaze se u medicinskim ili drugim zapisima

Vremenski izrazi

- Isti semantički tipovi vremenskih izraza
- Predloženo je uvođenje nove klase (PREPOSTEX)
- Proces normalizacije podrazumeva, ne samo utvrđivanje vrednosti iskazanog vremenskog izraza, već i njihovo pomeranje za određeni vremenski interval

Rezultat

- kreiranje proizvoda koji će, bez dodatne pripreme, omogućiti automatsko obeležavanje vremenskih izraza srpskog jezika
- novi izvori za dalja istraživanja u oblasti obrade prirodnojezičkih tekstova korišćenjem statističkih metoda, odnosno korpusi (opšteg tipa i iz domena medicine) obeleženi dovoljnim brojem primera

Dalji rad

- dopuna postojećih pravila za prepoznavanje i normalizaciju vremenskih izraza
- uključivanje preostalih atributa definisanih TimeML shemom (TimeID, AncorID, BeginPoint, EndPoint)
- automatska obrada događaja i vremenskih relacija

Hvala na pažnji!

jjacimovic@afrodita.rcub.bg.ac.rs