Automatsko konvertovanje pravnih propisa Republike Srbije u Akoma-Ntoso *v*3.0 format podataka

Andrija Cvejić
Računarstvo i automatika
Fakultet tehničkih nauka, Univerzitet u Novom Sadu
Novi Sad, Srbija
andrijacvejic@uns.ac.rs

Katarina-Glorija Grujić
Računarstvo i automatika
Fakultet tehničkih nauka, Univerzitet u Novom Sadu
Novi Sad, Srbija
katarina.glorija@uns.ac.rs

Aleksandar Cvejić
Računarstvo i automatika
Fakultet tehničkih nauka, Univerzitet u Novom Sadu
Novi Sad, Srbija
aleksandarcvejic@uns.ac.rs

***Abstrakt*—Akoma Ntoso (*eng. Architecture for Knowledge-Oriented Management of African Normative Texts using Open Standards and Ontologies*) je afrički standard koji omogućuje opisivanje pravnih dokumenata. Rad se bazira na transformaciji pisanog (eng. *plain*) teksta propisa Republike Srbije u *XML* (eng. *eXtensible Markup Language*). Za validaciju se koristi podšema Akoma Ntoso verzija 3.0. Propisi mogu biti zakoni, odluke, pravilnici, uredbe i drugi oblici pravnog propisa. Proces *XML* obeležavanja se sprovodi u tri sloja: metapodaci, hijerarhijsko-strukturalni sloj i tekstualno-sadržajni sloj. Za pronalaženje bitnih delova u dokumentima, sistem koristi *REGEX* (eng. *Regular expression*) šablone. Rezultat izvršavanja programa ovog rada jeste validan *XML* dokument pod 3.0 Akoma Ntoso šemom. Dokumenti dobijeni kao izlaz iz sistema su univerzalni i mogu biti korišćeni od strane bilo kog drugog sistema koji koristi ovaj standard. Na taj način dokumenti postaju mašinski čitljivi. Takođe, *XML* dokumentima, koji su u formi Akoma Ntoso notacije, lakše se upravlja u bilo kom koraku zakonodavnog ili sudskog životnog ciklusa.**

Ključne reči — Akoma Ntoso; *REGEX*; *XML*; *NER*; propisi; *tf-idf; crf*; zakon;

I. Uvod

Akoma Ntoso je standard za obeležavanje pravnih propisa razvijen uz podršku Ujedinjenih Nacija (UN/DESA), sa ciljem kreiranja standardnog načina za razmenu dokumenata između informacionih sistema Afrike. Međutim, danas upotreba standarda se proširila i na države drugih kontinenata kao što su Azija, Amerika i Evropa. Njegova široka upotreba je zahvalna time što dokumenti u Akoma Ntoso standardu su napravljeni od više odvojenih slojeva, kao što su metapodaci, strukturalni i tekstualni. Najbitnija karakteristika standarda je očuvanje originalnog teksta u tekstualnom sloju, gde se tekst rasporedi u određenu hijerarhijsku struktura elemenata (eng. *tags*). Sloj metapodataka sadrži dodatne informacije o dokumentu, koje su namanjene za semantičko obagaćenje informacija u dokumentu za mašine, gde se uljučuju reference ka elementima ontologija. Takođe, Akoma Ntoso je prikladan standard za čuvanje podataka kao model podataka dokumenata u sistemima. To se postiže pomoću *XML* šeme koja se može koristiti prilikom kreiranja, modifikacija i komunikacije informacionih sistema za pravne procese. Primer takvih sistema su informacioni sistemi vlade, skupštine, pokrajina i drugi. Rešenje u radu je inspirisano radom kolege[13].

Pravni propisi su preuzeti sa Pravno-informacionog sistema Republike Srbije[2] koji su većinom formalni, odnosno poštuju "Jedinstvena metodološka pravila za izradu propisa Republike Srbije"[9]. Među objavljenim propisima, veliki broj njih su u tekstualnom ili *HTML* formatu. Zbog toga, sistem je namenjen kao pomoćno sredstvo za već napisane pravne propise da se obeleže u Akoma Ntoso 3.0 notaciji.

Prilikom transformacije u Akoma Ntoso standard obeležavanja, potrebno je izdvojiti elemente pravnih propisa Republike Srbije koji su ekvivalentni elementima u specifikaciji Akoma Ntoso 3.0 verzija *XML* šeme.

Struktura rada je ogranizovana na sledeći način: Sekcija II opisuje postupak odabira i prilagođavanje *XML* elemenata iz Akoma Ntoso šeme. Sekcija III opisuje proces prikupljanja i prečišćavanja podataka, koji zatim predstavlja ulaz u ostatak sistema. Sekcija IV objašnjava postupak prepoznavanja metapodataka i način predstavljanja pomoću Akoma Ntoso šeme. Sekcija V se bavi preslikavanjem hijerarhije propisa u hijerarhijsku strukturu Akoma Ntoso. U sekciji VI predstavljen je tekstualni sloj i opisan postupak prepoznavanja i formiranja referenci. U sekcija VII izvršena je analiza dobijenih rezultata sistema. Sekcija VIII predstavlja zaključak ovog rada i iznosu moguće pravce daljeg istraživnja.

II. Prilagođvanje Akoma Ntoso šeme za srpske propise

Za poštovanje podšeme Akoma Ntoso 3.0 standarda[1], koja je prilagođena restrikcijom za srpske zakone, neophodno je ispoštovati dve stvari. Prvo, potrebno je izabrati hijerarhijske elemente iz standarda koji se preslikavaju na strukturu srpskih propisa. Drugo, potrebno je poštovati restrikcije prilagođene podšeme.

978-1-5386-1066-4/18/$31.00 ©2018 IEEE

A. Izabrani hijerarhijski elemenenti iz skupa dozvoljenih

U tabeli 1. je predloženo rešenje preslikavanja elemenata, koji odgovaraju hijerarhiji srpskih pravnih akata na Akoma Ntoso elemente. Zvezdica(*) kod podtačke stavljena je zbog izuzetka jer podtačka nije dobila imenovano obeležje već proširiv element *hcontainer*. Da bi se označila podtačka, neophodno je dodati obavezan atribut *name* sa izabranom vrednosti *subpoint*.

TABELA I. HIJERARHIJSKI ELEMENTI

NAZIV HIJERARHIJE U PROPISU	NAZIV ELEMENAT U AKOMA NTOSO STANDARDU
DEO	*PART*
GLAVA	*CHAPTER*
ODELJAK	*SECTION*
PODODELJAK	*SUBSECTION*
ČLAN	*ARTICLE*
STAV	*PARAGRAPH*
TAČKA	*POINT*
PODTAČKA	*HCONTAINER**
ALINEA	*ALINEA*

B. Restrikcije podšeme nad Akoma Ntoso 3.0

Ideja realizacije proširivanja šeme je zasnovana nad jednim od predloga od strane *OASIS* grupe. Usvojen predlog se odnosi na korišćenje generatora modula šeme Akoma Ntoso. Postoji veliki izbor takvih generatora na internetu[11], oni pružaju lakše ažuriranje verzija i manju mogućnost kolizije sa originalnom šemom. Izabran generator pruža izbacivanje nekorišćenih modula propisa kao što su debate, osude i drugi. Generator dozvoljava uvođenje restrikcija koje su uvedene za poštovanje hijerarhije izabranih elemenata za sprske propise i takođe dozvoljenih podelemenata u njima. Za restrikcije je korišćen element iz *XML* šema verzije 1.1 zvani *assert* sa atributom *test,* gde je vrednost atributa *test xpath* izraz *"every $x in (*) satisfies $x/name() = ('elements','elements')".* Svi elementi imaju mogućnost da sadrže neki od dozvoljenih podelemenata kao što su *num*, *heading*, *subheading*, *content*. U tabeli 2. su prikazani dodatni dozvoljeni izuzeci restrikcije mogućih podelemenata u elementu, odnosno vrednost *elements* u prethodnoj formuli.

TABELA II. RESTRIKCIJE NAD PODELEMENTIMA

Naziv elementa	**Dodatni dozvoljeni podelementi (deca)**
Part	*chapter*
Chapter	*section, article, intro*
Section	*subsection, article*
Subsection	*article*
Article	*paragraph, intro*
Paragraph	*point, alinea, intro*
Point	*hcontainer, intro*
Hcontainer	*alinea, intro*
ALINEA	*//*

Uz pomoć generisane podšeme Akoma Ntoso, moguća je validacija srpskih pravnih propisa dokumenata tokom prijema, nakon kreiranja, modifikacije i drugih akcija. Time obezbeđujemo stabilnost strukture dokumenata i lakše korišćenje u budućnosti.

III. PRIKUPLJANJE I PRETPROCESIRANJE PODATAKA

Podaci su preuzeti (eng. *scraping*) sa zvaničnog sajta Pravno-informacionog sistema Republike Srbije[2]. Preuzeto je 1130 zakona, 2667 odluka, 2933 pravilnika i 1388 uredbi. Dokumenti su preuzeti u *HTML* (*Hypertext Markup Language*) formatu. Preuzeti su i metapodaci o svakom dokumentu koji su takođe dostupni na sajtu. Metapodaci predstavljaju dodatne informacije koje su od velikog značaja za pravni akt.

S obzirom na to da je ideja projekta da se transformiše običan (eng. *plain*) tekst u *XML* dokument, neophodno je bilo pretprocesirati preuzete akte. Pretprocesiranje predstavlja "čišćenje" teksta tako da se izbace svi *HTML* elementi. Takođe, izbačeni su i svi *CSS* (*Cascading style sheets*) stilovi. Izvršena je transformacija (*encode*-ovanje) tekstualnog sadržaja svih propisa u *encode*-ovanu vrednost (npr. ">" prebačeno je u ">"). Slova koja su *encode*-ovana su *XML* osetljiva slova.

IV. IZDVAJANJE METAPODATAKA

Standard jasno razdvaja sadržaj i metapodatke, gde metapodaci mogu biti dodati nakon objavljivanja propisa. Oni predstavljaju koncept dodavanja "znanja na znanje", iako su te informacije jasno razumljive čoveku, većina informacija sadržana u njima se pretežno koristi prilikom računarskog pretraživanja[12]. Međutim, Akoma Ntoso format služi za opisivanje širokog spektra pravnih dokumenata, od sudskih presuda do zakona. Iz tog razloga treba izabrati odgovarajuće elemente koji će se koristiti za dostupne informacije pravnih propisa.

Dalje u tekstu biće rečeno nešto više informacija o izabranim podelementima elementa *meta* iz Akoma Ntoso šeme: *identification, publication, classification, lifecycle, workflow, references* i *notes.*

A. Identifikacija dokumenta

Identifikacioni element se opisuje pomoću *FRBR* (*Functional Requirements for Bibliographic Records*) modela. *FRBR* model je konceptualni standard za bibliografske zapise, napravljen za očuvanje veza između dokumenata i da oni budu jednoznačno prepoznatljivi. Da bi se to postiglo *FRBR* model sadrži četiri sloja. Svaki od njih predstavlja poseban nivo apstrakcije dokumenta [6]. Prva tri sloja su:

1. Delo (eng. *work*) – apstraktni koncept dokumenta, u našem slučaju pravni propisi (npr. zakon 3 iz 2005.)

2. Izraz (eng. *expression*) – posebna verzija akta koja se razlikuje po nekoj osnovi kao što su jezik ili verzija (npr. zakon 3 iz 2005. na mađarskom)
3. Manifestacija (eng. *manifestation*) – označava konkretan format podataka u kojem je dokument reprezentovan, za pravne propise najverovatniji format prikaza je format Akoma Ntoso, označen sa ekstenzijom *akn*. (npr. *PDF* verzija zakona 3 iz 2005 na mađarskom).

Poslednji sloj *FRBR* modela je stavka (eng. *item*), i predstavlja jedinstvenu fizičku manifestaciju dokumenta. Primer reference *item* bi se odnosilo na konkretnu fizičku kopiju dokumenta ili memorijsku lokaciju na računaru. Zbog takvih karakteristika element *item* nije korišćen uopšte. Na Figuri 1 je prikazan primer novokreiranog identifikaciong bloka.

```
<identification source="#somebody">
    <FRBRWork>
        <FRBRthis value="akn/rs/act/2009/36-3/main"/>
        <FRBRuri value="akn/rs/act/2009/36-3"/>
        <FRBRdate date="2009-01-01" name="Generation"/>
        <FRBRauthor as="#author" href="#somebody"/>
        <FRBRcountry value="rs"/>
    </FRBRWork>
    <FRBRExpression>
        <FRBRthis value="akn/rs/act/2009/36-3/srp@/main"/>
        <FRBRuri value="akn/rs/act/2009/36-3/srp@"/>
        <FRBRdate date="2009-01-01" name="Generation"/>
        <FRBRauthor as="#editor" href="#somebody"/>
        <FRBRlanguage language="srp"/>
    </FRBRExpression>
    <FRBRManifestation>
        <FRBRthis value="akn/rs/act/2009/36-3/srp@/main.xml"/>
        <FRBRuri value="akn/rs/act/2009/36-3/srp@.akn"/>
        <FRBRdate date="2009-01-01" name="Generation"/>
        <FRBRauthor as="#editor" href="#somebody"/>
        <FRBRformat value="xml"/>
    </FRBRManifestation>
</identification>
```

Figura 1. Izgled *FRBR* modela za zakon

Informacije koje su neophodne za formiranje identifikacionog bloka jesu datum usvajanja, verzija i jezik akta. Kôd jezika je napisan po *ISO* 639-2 standardu za obeležavanje jezika[8], stavlja se "srp" nezavisno da li je latinično ili ćirilično pismo u pitanju[7].

B. Publikacija, radni tok, klasifikacija i beleške dokumenta

Zajedno sa samim pravnim dokumentima koji su preuzeti sa sajta Pravno-informacionog sistema, za svaki dokument preuzeti su i metapodaci. Oni su predstavljeni u tabeli 3, zajedno sa njihovim odgovarajućim Akoma Ntoso elementima.

TABELA III. METAPODACI PRAVNIH DOKUMENATA

Element	Odgovarajuće informacije
publication	„Glasilo i datum objavljivanja"
workflow	„Datum stupanja na snagu osnovnog teksa", „Datum primene" i „Datum usvajanja"
classification	„Vrsta propisa", „Oblast" i „Grupa"
notes	„Napomena izdavača" i „Dodatne informacije"

C. Reference u metapodacima

Na nivou dokumenta se vrši prepoznavanje koncepata i entiteta (*eng. named entity recognition*). Prepoznate vrednosti generišu *TLC* (*eng. Top level class*) reference iz Akoma Ntoso neformalne ontologije. Tabela 4 sadrži sve *TLC* reference koje se pronalaze odgovarajućom tehnikom, nakon čega se generišu u rezultujućem fajlu.

TABELA IV. ZNAČENJE IZABRANIH *TLC* REFERENCI

Element	Značenje u dokumentu
TLCConcept	Predstavlja apstraktni pojam ili ideju.
TLCLocation	Predstavlja lokaciju, bila ona teritorijalna, istorijska, geografska ili geopolitička.
TLCOrganization	Predstavlja ime organizacije ili grupu ljudi koji se identifikuju sa određenim nazivom.
TLCPerson	Predstavlja naziv individue.
TLCEvent	Obuhvata vremenske intervale i datume.

1) Način prepoznavanja koncepata

Korišćen je *tf-idf* [3] (*term frequency–inverse document frequency*) algoritam za ekstrakciju najrelavantnijih pojmova na nivou dokumenta. Ulaz za statistički algoritam je vektor reči. Iz vektora su izbačene srpske reči koje ne doprinose značenju u rečenicama (eng. *stopwords*). Za sam algoritam se posmatra jedan član kao dokument iz koga se vadi frekvencija reči.

Matematička jednačina 1. predstavlja prvi korak algoritma, gde se računa učestalost reči - pojmova (*TF - term frequency*). Ukupan broj pojavljivanja jedne reči u članu se deli brojem reči u članu. Svaki član propisa ima svoju učestalost reči.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \tag{1}$$

Nakon toga se računa inverzna frekvencija pojavljivanja reči (*IDF - inverse data frequency*). Prvo se deli broj članova (*N*) brojem članova koji sadrži datu ***w*** reč. Inverzna frekvencija određuje težinu reči koje se retko pojavljuju.

$$idf(w) = log(\frac{N}{df_t}) \tag{2}$$

Konačan rezultat bitnosti reči se dobija množenjem prethodno dobijenih vrednosti.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{3}$$

Za ovaj rad izabrano je odbacivanje svih pojmova koji nisu dobili bar 0.1 *tf-idf* vrednost. Koncepti koji su preostali se dodaju pod *reference* oznakom u rezultujućem *XML* dokumentu. Implementacija razmatranog algoritma je odrađena od strane *Scikit-learn* biblioteke [10].

2) Način prepoznavanja entiteta

Pronalaženje ostalih *TLC* referenci se vrši pomoću *NER* (*named entity recognition*) algoritma, gde je korišćen standardni model *CRF* (*conditional random field*) za predikciju sekvenci labela. *CRF* je zastupnik

metode za nadgledano učenje, zbog čega je potreban skup labeliranih podataka na srpskom jeziku. Izabran je *SETimes.SR* [4] koji ispunjava uslove količine podataka i jezika. Skup podataka sadrži anotirane entitete prikazane na figuri 2 (korišćen je *IOB tagging* sistem prilikom anotiranja). Pored preuzetih podataka, dodati su ručno anotirani podaci koji predstavljaju datum.

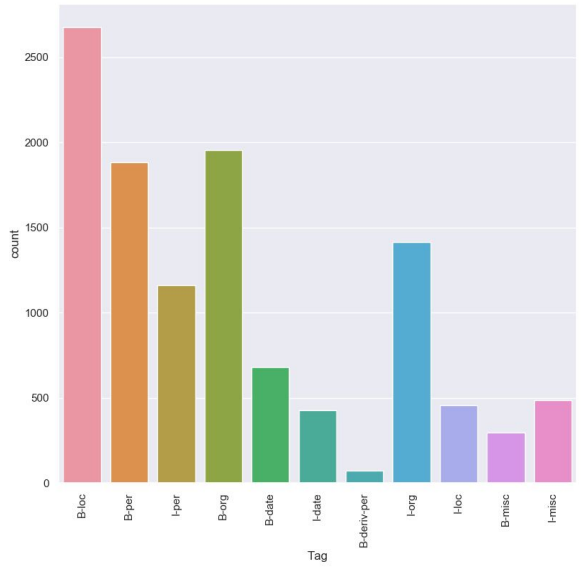

Figura 2. Raspodela anotiranih podataka izuzimajući "O" tag

Ulaz u predikcioni algoritam zahteva *POS* (part-of-speech) tagove zbog čega se koristi *ReLDI tokenizer* i *tagger* [5].

Za samu implementaciju izabran je *LGBFS* algoritam (*Gradient descent using the L-BFGS method*) unutar *Scikit-learn* biblioteke [10]. Na figuri 3 se vidi rezultat obučavanja nad izabranim podacima nakon validacije (korišćena je *5-cross-validation*).

	precision	recall	f1-score	support
B-date	0.93	0.93	0.93	681
B-deriv-per	0.84	0.89	0.86	75
B-loc	0.89	0.93	0.91	2678
B-misc	0.65	0.26	0.37	298
B-org	0.87	0.83	0.85	1953
B-per	0.90	0.92	0.91	1884
I-date	0.97	0.95	0.96	427
I-loc	0.82	0.76	0.79	459
I-misc	0.66	0.33	0.44	488
I-org	0.78	0.71	0.74	1416
I-per	0.91	0.96	0.94	1161
micro avg	0.87	0.84	0.86	11520
macro avg	0.84	0.77	0.79	11520
weighted avg	0.86	0.84	0.85	11520

Figura 3. Rezultat *CRF* obučavanja

Za optimizaciju hiperparametara *(c1,c2) CRF* modela korišćen je *Randomized CV Search*.

V. Hijerarhijska Struktura

Hijerarhijski (strukturni) sloj grupiše pravopisne klasifikacione celine u elemente. Za identifikaciju delova strukture u pravnim propisima korišćeni su regularni izrazi (eng. *Regular expression,* skraćeno *REGEX*). Regularni izrazi su napisani poštujući opis koji je dat za strukturu pravnih propisa, odnosno poštujući jedinstvena metodološka pravila za izradu propisa[9]. Prema pravilima o pisanju, pravni propisi su podeljeni u više strukturalnih celina, kao što su početak dokumenta (uvodni deo i preambula) i telo propisa.

A. Početak propisa (početak dokumenta)

Početak pravnog propisa može da sadrži preambulu, uvodni deo i naslov (izuzetak kod zakona je da oni nemaju preambulu)[9]. Takav sadržaj je neophodno prebaciti u odgovarajuće dostupne elemente Akoma Ntoso standarda. Sadržaj preambule je stavljen u element *preamble*, dok se u element *longTitle* smešta naslov propisa i gde se objavio. Sve ostale pronađene informacije se stavljaju u podelement *p* elementa *preface*. Na figuri ispod prikazan je izgled početka propisa nakon anotiranja.

```
<preface>
    <longTitle>
        <p>Уредба о утврђивању изворника Великог и Малог грба, изворника заставе и нотног
            записа химне Републике Србије "службени гласник рс", број
            <ref wId="ref0" href="akn/rs/act/2010/85/srp@">85 од 15. новембра 2010</ref>.</p>
    </longTitle>
</preface>
<preamble>
    <p>Влада доноси на основу <ref wId="ref1" href="akn/rs/act/09/36/srp@/!main~art_9__para_1" >
        члана 9. став 1. Закона о изгледу и употреби грба, заставе и химне Републике Србије
        („Службени гласник РС”, број 36/09)</ref> и
        <ref wId="ref2" href="akn/rs/act/08/65/srp@/!main~art_42__para_1"> члана 42. став 1.
        Закона о Влади („Службени гласник РС”, бр. 55/05, 71/05 – исправка, 101/07 и 65/08).</ref>
    </p>
</preamble>
```

Figura 4. Početak uredbe predstavljen u Akoma Ntoso standardu

B. Glavni deo propisa (Telo dokumenta)

Glavni deo propisa čine elementi u hijerarhijskoj (roditelj-dete) vezi. Moguće klasifiakcione jedinice poređane od najvećih ka najmanjim su: deo, glava, odeljak, pododeljak, član, stav, tačka, podtačka i alineja[9]. Osnovna klasifikaciona jedinica hijerarhije pravnog propisa je član, gde je neophodno da svaki broj člana bude jedinstven. To omogućuje da se navođenjem isključivo broja člana može pristupiti tekstu unutar njega. Zbog toga postoje logičke celine koje su veće od člana (deo, glava, odeljak, pododeljak) koje služe samo za smisleno odvajanje sadržaja. Zatim postoje manje klasifikacione jedinice (tačka, podtačka i alineja) koje služe kao strukturna podrška za bolje razlaganje člana. Svaka od spomenutih klasifikacionih jedinica ima propisan način označavanja u aktu[9]. Zbog regularnosti u podacima, možemo da vršimo ekstrakciju pomoću regularnih izraza.

Za svaku klasifikacionu jedinicu izabran je jedan element iz ponuđenih elemenata u Akoma Ntoso. Za svaku hijerarhijsku celinu propisa, napravljen je šablon ne bi li je prepoznavao. U tabeli 5 prikazane su vrednost *REGEX* šablona. Takođe, prikazan je i primer izgleda teksta koji se traži iz jednog pravnog propisa.

TABELA V. PRIKAZ HIJERARHIJSKIH ELEMENATA I NAČIN PREPOZNAVANJA

Hijerarhijski element	**Primer**	***Regex***
Deo	ПРВИ ДЕО	(.*) (ДЕО|DEO)
Glava	I. ОСНОВНЕ ОДРЕДБЕ	(Глава|Glava)?[MDCLXVI]+(\. |)(.*)
Odeljak	1. Значење израза	([0-9]+)(\.)(.*)
Podeodeljak	а) Овлашћења надлежног органа	([а-шђјљћџ]|[a-zđžćčš])(\))(.*)
Član	Члан 1.	(Члан|Član) ([0-9]+)(\.)
Stav	Овим законом уређују се статус...	\n
Tačka	3) странац је лице које није држављанин...	([0-9]+)(\))(.*)
Podtačka	(1) Општи подаци: лично име, ...	(\()([0-9]+)(\))(.*)
Alineja	- 1. августа 2004. године – до 530 евра;	(– ?\s?)(.*)

VI. TEKSTUALNI SLOJ

U okviru samog tekstualnog sloja se mogu pojaviti elementi koji služe za strukturiranje, isticanje informacija, ukazivanje na drugi deo teksta ili dokumenta. Za označavanje ovakvih delova postoje Akoma Ntoso elementi koji se poklapaju sa *HTML* elementima (tabele, paragrafi i reference).

A. Prepoznavanje i formiranje referenci

Prepoznavanje referenci je izvršeno pomoću *REGEX* šablona. U početku, šabloni su pisani na osnovu pravila iz člana 36, jedinstvenih metodoloških pravila za izradu propisa [9]. Međutim, uočilo se da postoji veliki broj slučajeva gde referenciranje ne poštuje ova pravila. Iz tog razloga, neophodno je bilo uočiti sve slučajeve pojavljivanja referenci. Pokušan je pristup koristeći tehnike mašinskog učenja, ali rezultati koji su dobijeni nisu bili zadovoljavajući. Razlog toga je velika varijabilnost oblika u kojima se može naći neka reč (promena po padežu, rodu ili broju, korišćenje skraćenica za reči, itd.). Iz tog razloga, u ovom radu odabrani su *REGEX* šabloni za implementaciju ovog dela sistema. Traženje šablona izvršeno je ručno, prolazeći kroz pravne akte i uočavajući različite vrste referenci. Šabloni su podeljeni u 4 grupe:

1. Referenciranje člana, stava ili tačke u okviru posmatranog pravnog akta
2. Referenciranje člana koji se nalazi u drugom pravnom aktu
3. Referenciranje Službenog lista (Službenog glasnika)
4. Referenciranje opsega članova, stava ili tačke

Problem na koji se naišlo prilikom formiranja šablona je nekonzistentnost u navođenju referenci. Primer je prikazan na figuri 5. Na primeru su prikazane varijacije navođenja opsega referenci. Neophodno je bilo kreirati više različitih šablona kako bi se obuhvatile sve varijacije.

čl. 23. do 26. ovog zakona.

čl. 196, 197. i 198. ovog zakona

čl. 161–164b ovog zakona,

Figura 5. Primer nekonzistentnog navođenja referenci

Reference su formirane poštujući zvaničan pravilnik Akoma Ntoso za formiranje *XML* dokumenata [1]. Opšti oblik reference dat je na figuri 6.

```
akn/<država>/act/<godina publikovanja u formatu YYYY-MM-DD ili samo YYYY>/
<broj akta u godini ili ako se ne zna "nn">/srb@/<!main, !imedoc.akn ili !schedule_1.pdf>/
art_<broj člana>__para_<broj stava>__point_<broj tačke>
```

Figura 6. Opšti oblik reference

Na figuri 7 prikazani su izgledi različitih tipova referenci. Prikazana su dva primera navođenja opsega referenci. Kao što se može primetiti, različito su formirane reference ukoliko se opseg navodi koristeći predloga "do", ili navođenjem svih elemenata. Prilikom formiranja referenci nemoguće je znati da li se između dva navedena elementa nalazi još neki (na primeru ispod, nemoguće je znati da li postoji član 41a). Ukoliko bi postojao, ta referenca ne bi predstavljala opseg i bilo bi pogrešno formirati je koristeći operator "->". Iz tog razloga, reference su formirane zasebno za svaki element. Ukoliko je referenca navedena koristeći predlog "do" ili znak "-", tada je očigledno u pitanju opseg, te je u tom slučaju referenca formirana koristeći operator "->".

```
„Службени гласник СРС", број <ref wId="ref13" href="akn/rs/act/85/6/srp@">6/85</ref>
<ref wId="ref152" href="akn/rs/act/2018/24-23/srp@/!main~art_87__para_1__point_6" >
                                                    члана 87. став 1. тачка 6</ref>)
<ref wId="ref13" href="akn/rs/act/2018/24-70/srp@/!main~art_7->art_11" >чл. 7. до 11. </ref>
чл. <ref wId="ref8" href="akn/rs/act/2009/36-3/srp@/!main~art_41" >41</ref>.
    и <ref wId="ref9" href="akn/rs/act/2009/36-3/srp@/!main~art_42" >42</ref>.
```

Figura 7. Primer različitih tipova referenci

VII. REZULTAT AUTOMATSKOG ANOTIRANJA

Dato poglavlje se bavi zapažanjima raličitih delova rešenja problema. Rešenje je razvijeno u *Python* programskom jeziku verzija 3.6.

A. Metapodaci

Za sve podelemente koji pripadaju elementu *meta* važi da su adektvatno kreirani. Ovi elementi mogu biti kreirani ukoliko su poznate informacije koje njima pripadaju. U slučaju da one nisu poznate stavlja se podrazumevana vrednost "neznanja". Izuzetak su rezultati elementa *references* sa *TLC* podelementima, koji će detaljnije biti analizirani u nastavku.

1) Rezultati dobijenih koncepata

Kvalitet koncepata zavisi od lematizacije i *tf-idf* algoritma. Prilikom izvlačenja koncepata, primećeno je da *tf-idf* algoritam daje adekvatne rezultate za slučajeve sa velikom varijacijom tema u članovima. Urađen je pregled na uzorku propisa. Rezultat analize na primerku od 50 propisa se može videti na figuri 8, gde se vidi da sam broj reči ne utiče direktno (nakon 1000 reči) na količinu ekstrakovanih koncepata. Na samoj figuri tamniji podeljci označavaju da ima više propisa sa tim brojem reči.

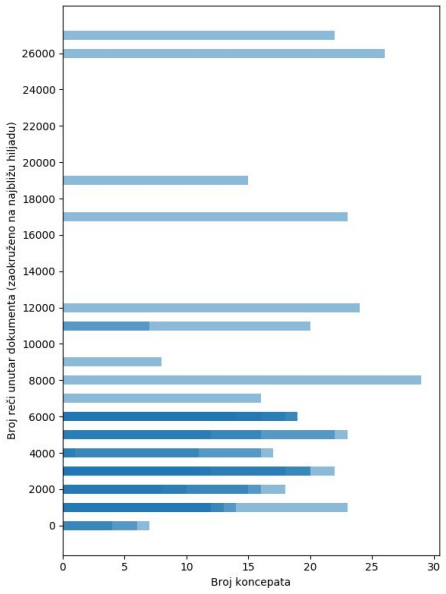

Figura 8. Analiza broj koncepata zavisno od broj reči u dokumentu

Za prethodni uzorak primećuje se prisustvo sledećih koncepata:

- zakon,
- član,
- stav i
- slični opšti pojmovi u zakonodavnom tekstu.

Ovi koncepti se mogu dodatno izbaciti, jer ne doprinose nekom dodatnom znanju. Na figuri 9, može se videti najfrekventnije pronađeni koncepti na prethonom uzorku.

Figura 9. Najfrekventniji konpceti dobijeni na uzorku

2) Rezultati dobijenih entiteta

Rezultat ranije diskutovan na figuri 3, iako tačan, prikazuje samo sliku nad podacima trening skupa. Korišćen korpus je preuzet iz svakodnevnih izvora, a ne iz pravnih propisa, zbog čega ne radi najbolje što bi algoritam mogao. Primećen je problem sa prepoznavanjem istih pojmova u različitim oblicima, što bi dodatno lematizacija mogla da pripomogne. Takođe je uviđen problem sa datumima, zbog toga što nisu prvobitno predviđeni svi oblici datuma u skupu podataka. Prilikom dodavanja datuma, anotirani su podaci u obliku:

- dd.mm.yyyy.
- mm.yyyy.
- *month*
- *dd. month*
- yyyy.

Unapređenje ovog rešenja podrazumevalo bi proširenje skupa podataka, balansiranje klasa i uvođenja kontraprimera koji bolje odgovaraju pravnim propisima.

B. Rezultati kod hijerarhije

Za raspoznavanje hijerarhijske strukture koristi se *reasoner* i *tokenizer* koji radi na *REGEX* principima. *Tokenizer* brine o tome koja klasifikaciona jedinica je u pitanju, dok *reasoner* odlučuje šta će se kreirati od dobijenog tokena. Uočeno je da dolazi do grešaka prilikom raspoznavanja hijerarhije invalidno napisanih akata, odnosno nepoštovanja metodoloških pravila za propise[9]. Uočeno je da nisu poštovana pravila iz tri razloga. U prvu grupu spadaju propisi koji su napisani pre unifikacije pravila pisanja propisa, tada su korišćeni razni stilovi. U drugu grupu spadaju propisi koji nisu u potpunosti ispoštovali pravila[9], to mogu biti male varijacije koje drastično utiču na rezultat, kao što su:

1. Nedostajuće tačke kod stavova. Tada nije lako razlikovati naslov u odnosu na stav, pošto stav nema jasnu oznaku.
2. Drugačije obeležavanje odeljaka i tačaka koji stvara konfuziju. Uočeno je pogrešno obeležavanje ovih struktura, gde postoji mogućnost da počinju sa “1)” ili “1.”. Ovaj problem je ispravljen proverom sledećeg tokena. Ako se ponavlja “token_odeljak” pretpostavi se da su tačke u pitanju.
3. Drugačije obeležavanje glava. U nekim primerim je uočeno obeležavanje “Glava prva” umesto “Glava I” ili da je izostavljena tačka na kraju. Ovaj slučaj je uzet u obzir u rešenju.

U treću grupu spadaju prevedeni propisi, odnosno propisi koji su napisani na stranom jeziku i prevedeni. Međutim nije prevedena struktura, već je ostavljena hijerarhijska struktura druge države. Takođe u treću grupu spadaju i propisi koji predstavljaju sporazume sa drugim balkanskim državama. Iako je sadržaj razumljiv, struktura je drugačija.

C. Ispravka grešaka kod reasoner-a

Iako postoji mogućnost greške kod automatskog anotiranja pravnog propisa, veliki broj njih se mogu automatski primetiti pomoću validacije šemom. Zbog

toga što svi dokumenti poštuju originalnu Akoma Ntoso 3.0 šemu, njena validacija nije od značaja. Iz tog razloga je napravljena Akoma Ntoso 3.0 šema za propise Republike Srbije. Rešenje je koristilo implementaciju *xmlschema python* biblioteke, jer je ona jedna od biblioteka koja podržava proveru *XML* šemom verzije 1.1 nad dokumentima.

D. Provera rezultata u odnosu na ručno anotirane

Za meru tačnosti rezultata automatsko anotiranih dokumenata pravnih propisa korišćene su F1 mera i sekvencijalna sličnost teksta. Za izračunavanje ovih metrika koriste se ručno anotirani dokumenti. F1 mera prikazuje realnije stanje o tačnosti sistema i ona se računa pomoću preciznosti (eng. *precision*) i povrata (eng. *recall*). Formule 3 i 4 prikazuju računanje preciznosti i povrata. Za forumule 3 i 4 neophodno je znati šta su *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) i *False Negative* (FN). Ukratko njihova značenja su:

1. TP-Broj podataka koji jesu tačani
2. TN-Broj podataka koji nisu tačani
3. FP-Broj podataka koji su netačni, a predstavljeni su kao tačni
4. FN-Broj podataka koji nedostaju, a zapravo su tačni

$$Precision = TP/(TP+FP) \quad (3)$$

$$Recall = TP/(TP+FP) \quad (4)$$

Formula 5 prikazuje računanje F1 mere. Ona je u rasponu od 0 do 1, gde 0 predstavlja loš rezultat i 1 predstavlja savršen.

$$F1 = 2*(Recall * Precision) / (Recall + Precision) \quad (5)$$

Dobijene F1 mere:

1. Za proveru validnosti kreiranih referenci unutar teksta dobijena *F1* mera iznosi 0.76.

   Analiza: Iako je obuhvaćen veliki broj izuzetaka koji ne poštuju pravila napisanih u metodologiji[9], nije bilo moguće obuhvatiti sve slučajeve zbog mogućnosti pojavljivanja kolizije. Nakon analiziranja propisa, odlučeno je da se u obzir uzmu izuzeci koji se često pojavljuju, dok se ostali izuzeci mogu ručno ispraviti od strane korisnika.
2. Za proveru validnosti kreirane hijerarhijske strukture klasifikacionih jedinica pravnih propisa *F1* mera iznosi 0.95.

   Analiza: Rezultat je takav zbog velike pokrivenosti. Pokrivena je većina slučajeva koja ne poštuju pravila napisana u metodologiji[9]. Zbog velikog broja pojavljivanja izuzetaka, *reasoner* podržava odlučivanje različitih mogućnosti pisanja. U nekim slučajevima proverava validnost ulaznog tokena na druge načine. Provera validnosti se vrši čak i u slučaju validno napisanog tokena (kao u tabeli 5).

Provera sekvencijalne sličnosti izlaznih dokumenata u poređenju sa ručno anotiranim dokumentima propisa je vršena pomoću sličnosti (eng. *similarity*) i iznosi 0.87.

Analiza: Rezultat je takav zbog postojanja redudantinih ponavljanja kod automatsko anotiranih podataka. Dok bi korisnik anotirao jedanput takvo pojavljivanje i sa punim kontekstom rečenice, u skupu *TLC* elemenata reference mogu da se odnose na istu reč u drugom padežu.

VIII. ZAKLJUČAK

U radu je predstavljen sistem za automatsko anotiranje *XML* dokumenata. Na osnovu dobijenih rezultata zaključeno je da veliku ulogu u tačnosti anotiranja igra poštovanje metodologije za pisanje pravnih propisa. Nažalost, zbog velikog broja propisa koji ga nepoštuju, ovaj sistem nije u mogućnosti da anotira tačno svaki dokument. Ipak, prilikom automatskog anotiranja jednostavno je uočiti greške koje nastaju validiranjem dokumenta *XML* šemom. Uočeno je da su greške minimalne i da se mogu rešiti od strane korisnika alata.

Iz tog razloga, ovaj alat je namenjen da bude pomoćni program (eng. *plug-in*). On bi ubrzao proces prevođenja pravnih propisa u mašinski čitljive dokumente, ali bi neophodno bilo prethodno proveriti njihovu validnost i ispraviti ukoliko je to neophodno.

Takođe, uočeno je da se algoritmi mašinskog učenja mogu koristiti u ekstrakciji koncepata i entiteta koji se pojavljuju u zakonu. Pored toga što njihovom ekstrakcijom dokument postaje mašinski čitljiviji, moguće je vršiti i pretraživanje na osnovu dobijenih podataka.

Prepreka dobre implementacije *NER* algoritma pomoću mašinskog učenja jeste dobar algoritam za transformaciju teksta i sekvenci (*text* i *sequence embedding*). U ovom radu je korišćen *feature dictionary,* koji sadrži *POS* tagove, informacije o samim rečima (da li je broj, da li počinje sa velikim slovom itd.) i informacije o okolnim rečima. U daljem istraživanju mogli bi se isprobati različiti algoritmi za *text* i *sequence embedding*. Potencijalni algoritmi mogu biti *GloVE (Global Vectors for Word Representation), Bert (Bidirectional Encoder Representations from Transformers)* ili *ELMo (Embeddings from Language Models)* za transformaciju teksta i *SGT (Sequence Graph Transform)* ili *PCA (Principal component analysis)* za transformaciju sekvenci.

Predlog algoritama koji mogu biti korišćeni u budućem rešenju za ekstrakciju koncepata su *df-icf*, *mtf-idf* ili *LDA* (*latent Dirichlet allocation*).

REFERENCE

[1] Akoma Ntoso version 3.0 XML schema (http://docs.oasis-open.org/legaldocml/akn-core/v1.0/os/part2-specs/schemas/akomantoso30.xsd)

[2] Pravno-informacioni sistem srbije (http://www.pravno-informacioni-sistem.rs/)

[3] Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2015). Research-paper recommender systems: a literature survey.

International Journal on Digital Libraries, 17(4), 305–338. doi: 10.1007/s00799-015-0156-0

[4] Batanović, Vuk; Ljubešić, Nikola; Samardžić, Tanja and Erjavec, Tomaž, 2018, Training corpus SETimes.SR 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1200.

[5] Ljubešić, N., Klubička, F., Agić, Ž., & Jazbec, I. P. (2016, May). New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4264-4270).

[6] Hickey T.B., O'Neill E.T., Toves J., 2002. Experiments with the IFLA functional requirements for bibliographic records (FRBR). *D-Lib magazine*, 8(9), pp.1-13.

[7] Vitali F., Palmirani M., & Parisse V. (2017, April). *Akoma Ntoso Naming Convention Version 1.0*. OASIS Committee Specification Draft 03 / Public Review Draft 03. http://docs.oasis-open.org/legaldocml/akn-nc/v1.0/csprd03/akn-nc-v1.0-csprd03.html

[8] Byrum, J. D. (1999). ISO 639-1 and ISO 639-2: International Standards for Language Codes. ISO 15924: International Standard for Names of Scripts.

[9] Zakonodavni odbor Narodne skupštine Republike Srbije, "Jedinstvena metodološka pravila za izradu propisa", *Službeni glasnik RS, br. 21/2010*, 30. marta 2010.

[10] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

[11] Korišćen modularan generator podšeme Akoma Ntoso verzije 3.0 http://akn.web.cs.unibo.it/akgener

[12] Barabucci, G., Cervone, L., Palmirani, M., Peroni, S., & Vitali, F. (2009, September). Multi-layer markup and ontological structures in Akoma Ntoso. In International Workshop on AI Approaches to the Complexity of Legal Systems (pp. 133-149). Springer, Berlin, Heidelberg.

[13] Dobrički, T. (2019) Sistem za automatsko konvertovanje pravnih akata Republike Srbije u Akoma-Ntoso format podataka, Fakultet tehničkih nauka, Novi Sad, Srbija