Softverski sistem za određivanje semantičke sličnosti kratkih tekstova na srpskom jeziku

Vuk Batanović, Bojan Furlan, Boško Nikolić

***Sadržaj* — U radu je opisan softverski sistem koji ocenjuje stepen semantičke sličnosti dva zadata kratka teksta na srpskom jeziku. Objašnjeni su osnovni principi na kojima sistem funkcioniše, kao i faze razvoja i evaluacije sistema. Takođe, opisan je postupak generisanja korpusa parafraza nad kojim je izvršena evaluacija. Na kraju, analizirani su rezultati evaluacije i razmotrene su mogućnosti poboljšanja preciznosti rada sistema.**

***Ključne reči* — similarity of short texts, corpus-based measures, paraphrase corpora construction.**

I. Uvod

SEMANTIČKA sličnost predstavlja koncept dodeljivanja metrike skupovima izraza ili dokumenata zasnovane na sličnosti njihovog značenja. Ovaj koncept je jedan od ključnih za razumevanje prirodnih jezika, jer omogućava pravljenje smislenih poređenja i zaključivanja. Zbog toga određivanje semantičke sličnosti igra važnu ulogu u automatskoj kategorizaciji i sumarizaciji tekstova, mašinskom prevođenju, pronalaženju informacija i drugim oblastima veštačke inteligencije. Semantičko poređenje kratkih tekstova ima poseban značaj, jer se kratki tekstovi, u vidu pitanja i rečeničnih izraza, danas masovno koriste kao upiti veb pretraživača. Pored toga, na Internetu je veliki deo vesti i informacija predstavljen kratkim tekstualnim navodima, te je pri automatskoj obradi takvih tekstova značaj određivanja semantičke sličnosti njihovih delova sve veći.

Statistička sličnost koristi vektorske prostore stanja da bi izrazila korelaciju reči dobijenih iz odgovarajućeg korpusa tekstova. Velika prednost statističkog pristupa je u tome što on ne zahteva prethodno postojanje nikakvih modela značenja reči, za čije su formiranje potrebni znatni ljudski i vremenski resursi. Stoga je sistem za određivanje semantičke sličnosti prikazan u ovom radu zasnovan na korišćenju statističke sličnosti.

Osnovni postupak koji se primenjuje pri izračunavanju statističke sličnosti je konstruisanje semantičkog prostora na osnovu distribucije reči unutar korpusa tekstova. U takvom prostoru svaka reč ima svoj kontekstni vektor, a semantička sličnost među rečima je zapravo predstavljena relacijom između tih vektora. Ovaj zaključak je posledica distribucionalne hipoteze po kojoj reči sa sličnim značenjima imaju tendenciju da se pojavljuju u sličnim kontekstima [1].

U narednim odeljcima prikazana je upotreba postojećih alata u cilju izgradnje realizovanog softverskog sistema. Nakon toga su opisane faze rada sistema i faze njegove evaluacije. Konačno, razmotreni su dobijeni rezultati kao i mogućnosti njihovog poboljšanja kroz različite pristupe za unapređenje sistema.

II. Korišćeni alati

U ovom radu korišćeni su i ranije razvijeni alati i postupci za rešavanje problema [2]. Korišćeno je postojeće rešenje za steming reči na srpskom, a primenjeni su i ranije predstavljeni algoritmi određivanja leksičke i semantičke sličnosti i način njihovog kombinovanja. Takođe, upotrebljene su neke ranije prezentovane osnovne ideje za izradu resursa za evaluaciju sistema [3].

Steming (stemming) predstavlja transformaciju kod koje može doći do uklanjanja sufiksa reči pri čemu se ne gubi osnovni semantički sadržaj. Ovaj postupak se može shvatiti i kao proces normalizacije u kojem se nekoliko morfoloških varijanti mapira u isti oblik. Steming tako smanjuje broj različitih reči jer se sve reči sa istom osnovom mapiraju u isti oblik. Npr, reči šuma, šumski i šumovit se sve prevode u oblik „šum“. Za steming u engleskom jeziku je razvijen veći broj različitih rešenja, od kojih je najpoznatiji Porterov stemer. Međutim, što se srpskog jezika tiče, ova oblast je tek u začetku. Zbog toga je u ovom radu korišćeno najbolje dostupno rešenje – stemer zasnovan na algoritamskom pristupu [4]. Tačnost ovog stemera je 81,83%.

Leksička sličnost se zasniva na analizi leksičkog poklapanja reči odnosno delova reči. Po uzoru na rad [5], za određivanje leksičke sličnosti u ovom radu se koriste tri varijante LCS (Longest Common Subsequence) algoritma uz odgovarajuće normalizacije, a zatim se uzima prosek njihovih ocena. To su: NLCS (Normalized Longest Common Subsequence), $MCLCS_1$ (Maximal Consecutive Longest Common Subsequence starting at character 1) i $MCLCS_N$ (Maximal Consecutive Longest Common Subsequence starting at character N).

Za semantičko poređenje iskorišćeni su gotovi algoritmi za procesiranje korpusa tekstova iz S-Space paketa [6], koji je deo open source Google Airhead projekta. Na taj način uključene su kako funkcije za procesiranje korpusa, tako i alati za njegovo pretprocesiranje i postprocesiranje.

Ovaj rad je delimično finansiran od strane Ministarstva nauke i prosvete Republike Srbije (projekti III44009, 44006 i 32047).

Vuk Batanović, Bojan Furlan, Boško Nikolić, Elektrotehnički fakultet Univerziteta u Beogradu, Bulevar kralja Aleksandra 73, 11120 Beograd, Srbija (e-mail: vukbatanovic@sbb.rs, bojan.furlan@etf.bg.ac.rs, nbosko@etf.bg.ac.rs).

Algoritmi za procesiranje korpusa se zasnivaju na korišćenju matrice zajedničkog pojavljivanja reči (co-occurrence matrix). U njoj svaka vrsta predstavlja jedinstvenu reč **w**, a svaka kolona **Fc** reprezentuje kontekst **c**. Ćelija, tj. element matrice **Fwc**, sadrži broj pojavljivanja reči **w** u datom kontekstu. Kontekst može biti dokument ili region neke druge reči, u zavisnosti od algoritma. U slučaju dokumenta, dimenzije vektora će odgovarati ukupnom broju dokumenata, dok u slučaju reči dimenzije vektora mogu, u najgorem slučaju, odgovarati ukupnom broju različitih reči koje se mogu naći u korpusu. Algoritmi koji su korišćeni u ovom radu su: COALS (Correlated Occurrence Analogue to Lexical Semantic) [7] i RI (Random Indexing) [1].

COALS je odabran, jer postiže veću preciznost od starijih algoritama kao što su LSA (Latent Semantic Analysis) i HAL (Hyperspace Analogue to Language), uz neznatno duže vreme obrade. Razmotren je i RI algoritam koji je, zbog svog specifičnog inkrementalnog rada, naročito pogodan za obradu velikih korpusa tekstova.

III. Realizacija sistema

Pribavljanje korpusa je podrazumevalo pronalaženje dovoljno velikog, javno dostupnog i besplatnog korpusa tekstova koji bi mogao da posluži kao osnova za kreiranje semantičkog prostora. Kao dobro rešenje pokazao se korpus apstrakta članaka sa Vikipedije na srpskom jeziku koji je dostupan u vidu XML fajla.

Korišćeni korpus članaka sa Vikipedije je napisan delimično na ćirilici, a delimično na latinici i kodiran je u UTF-8 formatu. To je predstavljalo problem zato što korišćeni stemer za srpski jezik prihvata kao ulaz samo reči napisane u specijalnom dual1 kodiranju kod koga se svaki dijakritik koduje kombinacijom dva nedijakritička slova. Iz ovog razloga, bilo je potrebno najpre napraviti konvertor koji će tekst korpusa članaka sa Vikipedije prevesti sa ćirilice i latinice na ovo specijalno kodiranje. Nakon konverzije, bilo je potrebno izdvojiti tekst od interesa iz celog korpusa. Naime, analizom strukture XML fajla koji sadrži korpus apstrakta članaka utvrđeno je da se tekst svakog apstrakta nalazi između XML tagova <abstract> i </abstract>, dok ostatak fajla čine informacije nebitne za kreiranje semantičkog prostora. Stoga je trebalo izvršiti ekstrakciju željenog teksta.

Pretprocesiranje ulaznog korpusa tekstova služi da se smanji ukupan broj različitih reči u korpusu, čime se smanjuju dimenzije kontekstnih vektora reči, a time i opterećenje računarskih resursa. U ovom radu, pretprocesiranje se vrši u dva koraka – čišćenje teksta i steming. Čišćenje teksta podrazumeva uklanjanje karaktera koji spadaju u druga pisma, uklanjanje brojeva i reči koje sadrže brojeve, uklanjanje interpunkcije, i izjednačavanje malih i velikih slova.

Procesiranje korpusa podrazumeva izbor željenog algoritma za kreiranje semantičkog prostora i zadavanje fajla koji sadrži tekst korpusa.

U postprocesiranju se ostvaruje redukcija dimenzija kontekstnih vektora, odnosno redukcija dimenzija co-occurrence matrice, čime se smanjuje kompleksnost izračunavanja prilikom određivanja sličnosti rečenica. Svaki algoritam odvojeno sprovodi postprocesiranje, a taj proces je enkapsuliran u samim algoritmima, koji su deo S-Space paketa. COALS algoritam za postprocesiranje koristi SVD (Singular Value Decomposition) algebarsku operaciju koja se zasniva na faktorizaciji i dekompoziciji matrice. Sa druge strane, kod RI algoritma uopšte nije potrebno korišćenje tehnika postprocesiranja.

Snimanje semantičkog prostora je neophodno da se on ne bi morao iznova kreirati pri svakom pokretanju programa. Čuvanje semantičkog prostora u vidu fajla nije praktično zbog loših performansi nasumičnog pristupa jednom delu ogromnog fajla. Zato se semantički prostor smešta u bazu podataka, čime se garantuje prihvatljiva brzina pristupa. Za bazu podataka je odabran mySQL. Baza sadrži dve tabele – jednu namenjenu semantičkom prostoru koji je dobijen korišćenjem COALS algoritma i drugu namenjenu semantičkom prostoru dobijenom korišćenjem RI algoritma. Obe tabele imaju identičnu strukturu i sadrže kolone za ključ, reč i za odgovarajući kontekstni vektor koji je predstavljen kao dugačak string.

Prilikom određivanja sličnosti dva teksta, najpre se određuje leksička sličnost, a zatim i semantička. Pod određivanjem sličnosti se podrazumeva izračunavanje ocene sličnosti tekstova u opsegu od 0 do 1, pri čemu 0 označava potpunu semantičku različitost, a 1 potpuno semantičko poklapanje. Semantička sličnost dve reči se određuje kosinusnim poređenjem njihovih kontekstnih vektora koji su pročitani iz baze podataka. Konačna procena sličnosti se dobija kombinovanjem ocena leksičke i semantičke sličnosti, pri čemu oba tipa sličnosti u konačnoj proceni nose istu težinu.

IV. Evaluacija sistema

Glavni problem u evaluaciji predstavlja određivanje vrednosti parametara rada sistema za koje on postiže maksimalnu preciznost, pri čemu se pod preciznošću podrazumeva stepen poklapanja ocena sistema sa ocenama sličnosti koje bi dao čovek. U tu svrhu je neophodno odrediti optimalan prag za ocene sličnosti koje vraća sistem, tj. odrediti optimalnu vrednost između 0 i 1 koja bi predstavljala granicu, tako da se sve ocene iznad nje mogu tretirati kao procena semantičke sličnosti, a sve ocene ispod nje kao procena semantičke različitosti. Da bi se ovaj prag odredio, potreban je veliki skup parova rečenica koje su semantički bliske, od kojih neke zaista predstavljaju parafraze, a neke ne. Nakon što se ovakav korpus parova rečenica ručno oceni binarnim ocenama, moguće je poređenjem ručno dodeljenih i mašinski određenih ocena, statističkom analizom, odrediti vrednost praga za koju sistem postiže optimalnu preciznost.

Osnovni pristup za izgradnju korpusa se zasniva na pronalaženju više novinskih izveštaja koji se bave istom vešću. Po novinarskoj konvenciji, prva rečenica izveštaja, ili prve dve rečenice izveštaja obično predstavljaju sumarizaciju sadržaja vesti, te su te rečenice iz različitih izveštaja o istoj vesti dobri kandidati za postojanje

parafraznog odnosa. Glavni zahtev pri izgradnji korpusa vesti se odnosio na postojanje besplatno dostupne i sređene veb arhive vesti, koja omogućava lak pristup važnim vestima za željeni datum. Kao najbolje rešenje pokazao se sajt www.vesti.rs. Ovaj sajt predstavlja agregator vesti koji objedinjuje vesti svih većih medijskih kuća u Srbiji, kako televizijskih stanica i štampanih medija, tako i mnogih Internet magazina i portala. Ukupno se koristi preko 210 različitih izvora vesti. Odlučeno je da se za skupljanje parafraza koriste samo najvažnije vesti za svaki datum, jer je za njih najveća verovatnoća postojanja izveštaja iz više izvora. Sem toga, na taj način je moguće pribaviti parove rečenica koji se tiču raznih oblasti života i raznih mesta i aktera dešavanja, čime se sprečava opasnost od fokusiranosti korpusa parafraza na neku specifičnu temu. Da bi se obezbedilo dovoljno materijala za izgradnju korpusa, obrađene su najvažnije vesti iz cele 2010. godine i iz prvih sedam meseci 2011. godine.

Nakon što se dobije tekst svih izveštaja jedne vesti, potrebno je taj tekst prečistiti, podeliti na rečenice i proceniti kvalitet tih rečenica. Parsiranje i čišćenje tekstova vesti je vrlo problematičan zadatak zbog potpuno slobodnog formata u kome se tekstovi vesti pojavljuju, upotrebe tačke na mestima koja ne predstavljaju kraj rečenice, i raznih varijanti nepotrebnih informacija koje je potrebno ukloniti. Takve su, na primer, informacije o mestu dešavanja, vremenu dešavanja, izvoru vesti, i sl. Svakom paru rečenica koji ispunjava određene minimalne kriterijume u pogledu dužine rečenica i broja semantički bitnih reči u njima, pridružuju se atributi koji opisuju taj par i koji se koriste prilikom određivanja najboljeg tj. najkvalitetnijeg para rečenica za svaki članak. Ti atributi su: broj dugačkih reči u kraćoj rečenici, broj dugačkih reči u dužoj rečenici i broj dugačkih reči koje se javljaju u obe rečenice, bez uzimanja u obzir ponavljanja reči. Pod dugačkim rečima podrazumevaju se reči od bar šest slova, tj. one reči za koje je izvesno da su semantički bitne.

Od svih parova rečenica pridruženih jednom članku, potrebno je odrediti jedan za koji je najverovatnije da je najkvalitetniji. Za najbolji ili najkvalitetniji par može se smatrati onaj za koji je verovatno da njegove rečenice sadrže istu semantičku informaciju, i to rečenu na zaista drugačiji način. One rečenice za koje je malo verovatno da su semantički iste, ili one koje jesu semantički iste, ali samo zbog velike leksičke sličnosti među njima, mogu se smatrati lošim kandidatima. Dakle, glavni cilj pri izgradnji korpusa parafraza je postići dovoljno veliki procenat zastupljenosti zaista semantički istih parova rečenica, izbegavajući pri tome koliko je god moguće proste primere sličnosti koji proizlaze iz leksičkog poklapanja. Da bi se selektovale najbolje parafraze, potrebno je dodeliti im numeričku ocenu koja se izvodi na osnovu navedenih atributa svakog para rečenica.

Zaključeno je da je najbolje favorizovati rečenice slične dužine koje imaju oko 50% istih reči. Naime, kod parova rečenica dosta različitih dužina smanjuje se verovatnoća da su zaista u pitanju parafraze. Pokazuje se da kada je procenat sličnih reči visok, tada su najverovatnije u pitanju dve iste rečenice od kojih je jedna proširena nekom semantički nebitnom informacijom. S druge strane, kada je procenat istih reči nizak, tada je najverovatnije da je u pitanju neki semantički različit deo iste vesti. Pored toga, dodatna težina se daje kratkim parovima rečenica, jer kod njih svaka reč nosi proporcionalno veću težinu u kreiranju konačne ocene.

Prilikom ručnog ocenjivanja korpusa parafraza sprovođeno je i ispravljanje slovnih i drugih grešaka koje su posledica nesavršenosti izvornog teksta. Konačno, dobijen je korpus od 1194 para rečenica. Od toga, 553 para su ocenjena kao semantički podudarna, a 641 par kao semantički različit. Procentualno, semantički podudarnih parova ima 46,31%, a semantički različitih 53,69%.

Određivanje vrednosti praga se prvo vrši na većem skupu podataka tj. parova rečenica, koji se naziva skup podataka za treniranje (Train set). Tako dobijena vrednost praga se proverava na nezavisnom skupu podataka za testiranje (Test set) i, ako se i na njemu postigne zadovoljavajuća preciznost, ta vrednost praga se usvaja kao optimalna. Trening korpus čini 835 parova rečenica, a test korpus 359 parova.

V. Dobijeni rezultati

Prosečna vrednost ocena svih parova rečenica iz kreiranog korpusa parafraza je 0,59 pri korišćenju COALS algoritma, odnosno 0,65 pri korišćenju RI algoritma. Prosečna vrednost ocena semantički podudarnih parova iz korpusa je 0,67 za COALS, odnosno 0,71 za RI, a semantički različitih 0,52 za COALS, odnosno 0,59 za RI. Različite preciznosti pri ocenjivanju Train skupa podataka korišćenjem COALS i RI algoritma za različite vrednosti praga su prikazane u tabelama 1 i 2.

Prag	**Pravilno ocenjene rečenice pomoću COALS algoritma**		
	Semantički podudarne	Semantički različite	Ukupno
0,1	100%	0%	46,23%
0,2	100%	0%	46,23%
0,3	100%	0,22%	46,35%
0,4	99,22%	10,91%	51,74%
0,5	91,45%	46,77%	67,42%
0,593	74,09%	76,61%	**75,45%**
0,6	70,47%	77,5%	74,25%
0,7	39,9%	93,99%	68,98%
0,8	16,58%	99,55%	61,2%
0,9	1,55%	100%	54,49%

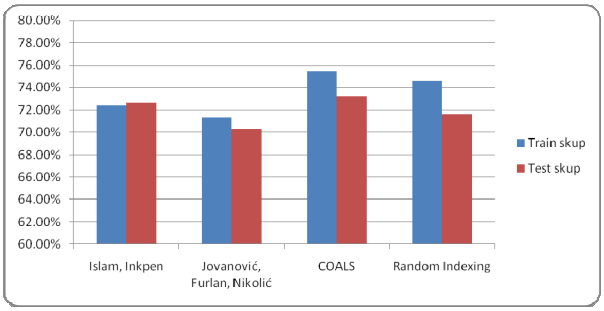
Tabela 1. Pregled ocena parova rečenica iz Train skupa dobijenih korišćenjem COALS algoritma

Najveći ukupan procenat tačno prepoznatih rečenica, tj. najveća preciznost sistema dobija se korišćenjem COALS algoritma za vrednost praga od 0,593, i tada je preciznost 75,45%, odnosno tada je pravilno ocenjeno 630 od ukupno 835 parova rečenica iz Train skupa. Proverom na Test skupu podataka dobija se preciznost od 73,26% za navedenu vrednost praga, koja se stoga usvaja kao optimalna. RI za optimalnu vrednost praga od 0,636 postiže preciznost od 71,59% na Test skupu podataka.

Na slici 1 predstavljeno je poređenje preciznosti sistema realizovanih u radovima [2] i [5] koji su se bavili izradom sistema za procenu sličnosti tekstova na engleskom jeziku, kao i sistema realizovanog u ovom radu. Kao što se vidi, preciznost sistema postignuta u ovom radu je nešto bolja od one koja je ostvarena u ranije implementiranim sistemima. Međutim, treba imati u vidu da je evaluacija sistema koji rade sa tekstovima na engleskom jeziku rađena uz pomoć MSRPC korpusa parafraza, koji je i veći od korpusa konstruisanog u ovom radu, i kvalitetniji, jer su postojala tri ocenjivača, te su ocene parafraza u tom korpusu pouzdanije.

Prag	**Pravilno ocenjene rečenice pomoću RI algoritma**		
	Semantički podudarne	Semantički različite	Ukupno
0,1	100%	0%	46,23%
0,2	100%	0%	46,23%
0,3	100%	0%	46,23%
0,4	100%	0,22%	46,35%
0,5	98,44%	12,03%	51,98%
0,6	87,82%	57,46%	71,48%
0,636	78,5%	71,27%	**74,61%**
0,7	55,44%	86,19%	71,98%
0,8	20,72%	99,11%	62,87%
0,9	3,11%	100%	55,21%

Tabela 2. Pregled ocena parova rečenica iz Train skupa dobijenih korišćenjem RI algoritma



Slika 1. Poređenje preciznosti sistema

Poređenjem vrednosti optimalnih pragova dobijenih u ovim radovima, vidi se da se one uglavnom kreću veoma blizu 0,6. Jedino se RI algoritam ističe nešto većom vrednošću optimalnog praga. Zanimljivo je primetiti da je vrednost praga za COALS algoritam, i vrednost usvojena u radu [2] u kome je takođe korišćen COALS, ali za tekstove na engleskom jeziku, skoro identična. Ako se uzme u obzir da je u radu [2] semantički prostor konstruisan na osnovu korpusa članaka sa Vikipedije na engleskom jeziku, koji je značajno veći od ovde korišćenog korpusa članaka na srpskom, kao i razlike između dva rada u pogledu veličine i kvaliteta korpusa parafraza koji su korišćeni za evaluaciju, može se zaključiti da se kod COALS algoritma vrednost optimalnog praga ne menja mnogo i u situacijama kada se radi sa resursima smanjenog obima i kvaliteta.

VI. Zaključak

U radu je opisan jedan pristup za određivanje semantičke sličnosti kratkih tekstova na srpskom jeziku. Opisana je realizacija sistema, kao i evaluacija sistema i njegovih rezultata.

Poboljšanje preciznosti realizovanog sistema se može ostvariti korišćenjem većeg korpusa za izgradnju semantičkog prostora. Što je veći korpus, to je i broj reči veći, pa je time veća i preciznost co-occurrence matrice. Drugi pristup bi predstavljalo unapređenje tehnika pretprocesiranja, tj. poboljšanje preciznosti stemera i implementacija izbacivanja stop-reči. Izbacivanjem stop-reči, postiglo bi se i smanjenje veličine co-occurrence matrice i povećanje preciznosti, jer bi tada matrica, rasterećena od semantički nebitnih reči, bolje odslikavala odnose između semantički bitnih reči.

Kvalitet evaluacije sistema bi se mogao poboljšati kako povećanjem broja parova rečenica u korpusu parafraza, tako i uvođenjem dodatnih procenjivača koji bi ocenili sve parove rečenica, čime bi se popravila pouzdanost ocena parafraza.

Literatura

[1] Magnus Sahlgren, An Introduction to Random Indexing, Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, 16 August 2005, Copenhagen, Denmark.

[2] Davor Jovanović, Bojan Furlan, Boško Nikolić, Softverski sistem za automatsko određivanje semantičke sličnosti kratkog teksta, 55. konferencija Društva za elektroniku, telekomunikacije, računarstvo, automatiku i nuklearnu tehniku ETRAN, jun 2011.

[3] Bill Dolan, Chris Quirk, Chris Brockett, Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources, Proceedings of the 20th International Conference on Computational Linguistics, August 2004.

[4] Vlado Kešelj, Danko Šipka, Pristup izgradnji stemera i lematizatora za jezike s bogatom fleksijom i oskudnim resursima zasnovan na obuhvatanju sufiksa, INFOTEKA – časopis za bibliotekarstvo i informatiku, broj 1-2, god. IX, maj 2008.

[5] Aminul Islam, Diana Inkpen, Semantic text similarity using corpus-based word similarity and string similarity, ACM Transactions on Knowledge Discovery from Data, Volume 2, Issue 2, July 2008.

[6] David Jurgens, Keith Stevens, The S-Space Package: An Open Source Package for Word Space Models, Proceedings of the ACL 2010 System Demonstrations, Uppsala, Sweden, 13 July 2010.

[7] Douglas L. T. Rohde, Laura M. Gonnerman, David C. Plaut, An Improved Method for Deriving Word Meaning from Lexical Co-Occurrence, Cognitive Science, March 22, 2004.

Abstract

This paper describes a software system for determining the degree of semantic similarity of two short texts written in Serbian. Its basic working principles are presented, as well as the phases of its functioning and evaluation. It also describes the process of generating a paraphrase corpus, which was used for evaluation purposes. Finally, evaluation results are discussed and further improvements of the system's precision are considered.

A SOFTWARE SYSTEM FOR DETERMINING THE SEMANTIC SIMILARITY OF SHORT TEXTS IN SERBIAN

Vuk Batanović, Bojan Furlan, Boško Nikolić