Evaluacija i klasifikacija korišćenja sintaksnih informacija u određivanju semantičke sličnosti kratkih tekstova

Vuk Batanović, Dragan Bojić

***Sadržaj* — U ovom radu su prikazani i kategorizovani načini korišćenja sintaksnih informacija u više algoritama za određivanje semantičke sličnosti kratkih tekstova. Evaluacija performansi algoritama je sprovedena korišćenjem rezultata testa detekcije parafraza iz Microsoft Research Paraphrase korpusa. Od svih opisanih algoritama i pristupa korišćenju sintaksnih informacija identifikovani su oni najpogodniji za primenu u jezicima sa ograničenim elektronskim jezičkim alatima i, imajući tu svrhu u vidu, predložena je nova klasifikacija algoritama.**

***Ključne reči* — semantic similarity, similarity of short texts, syntax information.**

I. Uvod

ODREĐIVANJE semantičke sličnosti tekstova zasniva se na dodeljivanju određene metrike paru tekstova na osnovu stepena povezanosti njihovih značenja. Sistemi za određivanje semantičke sličnosti kao izlaz daju ocene u opsegu od 0 do 1, gde 0 označava potpunu semantičku različitost zadatih tekstova, a 1 potpuno semantičko poklapanje. Semantička sličnost kratkih tekstova je posebno značajna jer su kratki tekstovi danas u širokoj upotrebi u vidu upita i rezultata pretraživača, naslova i sažetaka vesti, komentara na društvenim mrežama itd.

Veći broj problema u okviru oblasti procesiranja prirodnih jezika zavisi od upotrebe neke mere semantičke sličnosti tekstova. U takve probleme spadaju sumarizacija i kategorizacija tekstova, odgovaranje na pitanja, dohvatanje informacija, itd. Prilikom sumarizacije ključan korak predstavlja izbor rečenica koje će biti uključene u konačan tekst. U tom koraku važno je izbeći odabiranje rečenica koje sadrže iste informacije kao neka od već odabranih rečenica [1]. U sistemima za dohvatanje informacija ili odgovaranje na pitanja informacije iz upita mogu biti formulisane na drugačiji način od onoga korišćenog u dokumentu koji sadrži odgovor. Uzimanje tih varijacija u obzir može znatno poboljšati performanse sistema [2].

Postoje dva glavna pristupa određivanju semantičke sličnosti: *statistički*, zasnovan na korišćenju korpusa tekstova, i *topološki*, zasnovan na ekspertskom znanju. Statistički pristup je utemeljen na tzv. distribucionalnoj hipotezi po kojoj se reči sa sličnim značenjem javljaju u sličnim kontekstima [3]. Primenom te hipoteze na veliki korpus tekstova moguće je kreirati semantički prostor u kome je za svaku reč koja se javlja u korpusu specificirano koliko se puta javila u kakvim kontekstima. Pri tome, pod „kontekstom“ se obično podrazumeva ili dokument iz korpusa u kome se posmatrana reč javila ili neka druga reč u čijoj blizini se posmatrana reč javila. Svakoj reči se stoga dodeljuje odgovarajući kontekstni vektor što omogućava da se poređenje značenja reči svede na poređenje njihovih kontekstnih vektora. Prednost statističkog pristupa leži u tome što je za kreiranje ovakvog semantičkog prostora potreban jedino korpus tekstova na posmatranom jeziku.

U topološkom pristupu stepen semantičke povezanosti dve reči se određuje korišćenjem ručno modelovanih struktura podataka, kao što je *WordNet* za engleski jezik. Pošto su ove strukture konstruisane korišćenjem ljudskog, ekspertskog znanja, one mogu da dosta dobro modeluju stepene semantičke povezanosti među rečima. Njihova glavna mana leži u tome što su za njihovu izradu potrebni znatni ljudski i vremenski resursi.

U nastavku ovog rada dat je pregled različitih sintaksnih informacija i alata za njihovo dobijanje koje sistemi za određivanje semantičke sličnosti tekstova mogu da koriste. Nakon toga su prezentovani algoritmi iz ove oblasti i način na koji svaki od njih koristi sintaksne informacije. Zatim je izložena evaluacija svih pomenutih sistema i njihovih karakteristika i predložena nova vrsta njihove klasifikacije na osnovu mogućnosti njihove primene na jezike sa slabije razvijenim jezičkim alatima. Konačno, razmotreni su mogući pravci daljih istraživanja.

II. Sintaksne informacije i alati za njihovo dobijanje

Najosnovnija vrsta sintaksne informacije koja se može dobiti iz nekog teksta je redosled reči u tom tekstu. Za korišćenje redosleda reči u tekstu nisu potrebni nikakvi alati specifični za određen jezik, što čini ovaj tip informacija lako dostupnim u svim situacijama.

Najčešće korišćen tip sintaksnih informacija su podaci o vrsti reči – da li se radi o imenici, glagolu, pridevu, itd. Ovi podaci se dobijaju korišćenjem *part-of-speech* (POS) tagera koji su specifični za svaki jezik. Na primer, tageri za

Ovaj rad je deo istraživanja u okviru projekta podržanog od strane Ministarstva prosvete, nauke i tehnološkog razvoja TR32047.

Vuk Batanović, Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11120 Beograd, Srbija (e-mail: vukbatanovic@sbb.rs).

Dragan Bojić, Elektrotehnički fakultet, Univerzitet u Beogradu, Bulevar kralja Aleksandra 73, 11120 Beograd, Srbija (telefon: 381-11-3218-346, e-mail: bojic@etf.rs).

engleski su standardno u stanju da razaznaju 36 različitih vrsta i podvrsta reči definisanih u Penn Treebank projektu [4]. Određivanje vrste reči je problem klasifikacije, te se POS tageri obično realizuju primenom nadgledanog mašinskog učenja nad korpusom teksta koji je ručno anotiran tačnim POS tagovima. Savremeni POS tageri postižu veoma visoke nivoe tačnosti u radu od oko 97% [5] i predstavljaju važan jezički alat jer sve naprednije tehnike sintaksne analize zavise od njihovih rezultata.

Plitko parsiranje ili grupisanje (*chunking*) je operacija analize strukture rečenice koja identifikuje rečenične konstituente kao što su glagoli i sintagme. Plitko parsiranje ne specificira izgled unutrašnje strukture tih konstituenata, niti njihove uloge u rečenici. Nasuprot tome, pravi parseri formiraju stablo parsiranja u kome su reprezentovani i međusobni odnosi svih reči u rečenici i konstituenata koje te reči čine. Slično POS tagovanju, većina savremenih parsera su statističkog tipa, tj. zasnivaju se na upotrebi mašinskog učenja nad trening korpusom podataka koji je ručno parsiran. Stoga se parseri za svaki jezik moraju konstruisati zasebno, što je još teži postupak od kreiranja POS tagera, jer je konzistentno ručno parsiranje korpusa rečenica dugotrajan i naporan proces. Najsavremeniji statistički parseri dostižu visoke nivoe tačnosti od oko 91% [6], ali ipak imaju veći procenat grešaka od POS tagera.

Najnaprednija metoda korišćenja sintaksnih informacija je proces sintaksno-semantičke analize koji se naziva označavanje semantičkih uloga (SRL – *Semantic Role Labeling*). SRL predstavlja kompleksniju vrstu parsiranja koja je u stanju da označi koji rečenični konstituenti igraju koju ulogu u rečenici. Drugim rečima, SRL moduli su u stanju da prepoznaju koje reči ili grupe reči čine subjekat, koje predikat, koje objekat, itd. Njihova privlačnost leži u tome što daju informacije veoma visokog nivoa. Ipak, tačnost im je znatno manja od tačnosti klasičnih parsera i kreće se i u najboljim situacijama ispod 80% [7].

III. Algoritmi statističke sličnosti

Mihalcea et al. [8] su predložili nekoliko algoritama zasnovanih na istom osnovnom pristupu za određivanje semantičke sličnosti među kojima su dva koja koriste tehnike statističke sličnosti – PMI (*Pointwise Mutual Information*) i LSA (*Latent Semantic Analysis*). Za PMI je kao korpus korišćen sav sadržaj iz indeksa AltaVista pretraživača, dok LSA koristi Britanski nacionalni korpus koji sadrži oko 100 miliona reči. Pristup korišćen u ovom radu poredi tekstove na bag-of-words principu i od sintaksnih informacija koristi POS tagove radi sprečavanja uparivanja reči koje ne pripadaju istim vrstama reči.

Islam i Inkpen [9] su takođe zasnovali svoj pristup na bag-of-words principu, ali su koristili leksičku sličnost reči i drugačiju varijantu PMI algoritma nazvanu SOC-PMI (*Second Order Co-occurrence PMI*) koja je primenjena nad Britanskim nacionalnim korpusom. Furlan et al. [10] su predložili izmenjenu verziju Islam i Inkpen pristupa koja umesto SOC-PMI koristi COALS (*Correlated Occurrence Analogue to Lexical Semantics*) algoritam.

Lintean i Rus [11] su u okviru bag-of-words pristupa primenili LSA algoritam na TASA korpus od oko 16 miliona reči. U njihovom radu se, nalik na [8], koriste POS tagovi radi sprečavanja mogućnosti da budu uparene reči koje ne pripadaju istim vrstama reči. Za razliku od [8], Lintean i Rus razmatraju i relaksiraniju verziju ograničenja u kojoj se restrikcije vrše korišćenjem osnovnih vrsta reči, kao što su glagoli, a ne njihovih sitnijih kategorija kao što su infinitivi, prezenti, prošla vremena, itd.

Blacoe i Lapata [12] su predložili model distribuirane memorije (DM) koji je zasnovan na tenzorskom računu i za čije formiranje je korišćen POS tager i parser. Za izgradnju ovog modela je upotrebljeno nekoliko velikih korpusa, ukupne veličine preko 3 milijarde reči, među kojima su Britanski nacionalni korpus i celokupan sadržaj Vikipedije na engleskom jeziku. Pored ovog, razmotren je i jednostavniji SDS (*Simple Distributional Semantic Space*) pristup koji je zasnovan na direktnoj primeni distribucionalne semantike nad Britanskim nacionalnim korpusom. Za razliku od DM modela, SDS model ne koristi sintaksne informacije niti tehnike pretprocesiranja.

IV. Algoritmi topološke sličnosti

Mihalcea et al. [8] su, koristeći isti pristup primenjen na algoritme statističke sličnosti, razmotrili i šest algoritama koji koriste WordNet kao bazu znanja. Ti algoritmi se međusobno razlikuju po metrikama koje primenjuju za određivanje semantičke distance između dve zadate reči.

Fernando i Stevenson [13] su predložili bag-of-words model na koji je primenjeno šest WordNet metrika. Pošto je većina metrika ograničena na poređenje značenja samo između reči koje pripadaju istim vrstama reči, korišćen je POS tager na način sličan onome primenjenom u [8].

Li et al. [14] su zasnovali svoj pristup na plitkom parsiranju pomoću koga razlažu svaku rečenicu na glagole i imeničke, pridevske i priloške sintagme. Poređenje značenja rečenica se izvodi poređenjem odgovarajućih rečeničnih konstituenata. Značenja individualnih reči Li et al. dobijaju iz WordNet-a, ali koriste i Britanski nacionalni korpus radi dobijanja učestanosti pojavljivanja svake reči, koja takođe ulazi u finalnu ocenu sličnosti.

Furlan et al. [10] u svom radu predlažu i jedan algoritam zasnovan na korišćenju semantičke mreže *ConceptNet* kao baze znanja. Oni koriste SRL mehanizam koji je u stanju da iz rečenice izdvoji subjekat, predikat i objekat.

Kao i [13], i Oliva et al. [15] razmatraju više različitih WordNet metrika za određivanje semantičke sličnosti dve reči i koriste POS tager iz istog razloga kao i [8], [13]. Međutim, model koji Oliva et al. predlažu je dosta složeniji jer se u njemu, korišćenjem parsera i SRL modula, vrši detaljna i duboka sintaksna i sintaksno-semantička analiza rečenica koje se razmatraju. Kada završi obradu sintaksnih informacija njihov model vrši semantičko poređenje reči koje imaju istu sintaksnu funkciju iz obe rečenice. Model takođe penalizuje finalne ocene sličnosti ako jedna rečenica sadrži neke sintaksne strukture koje se ne pojavljuju u drugoj rečenici, npr. ako se objekat javlja samo u jednoj, a ne i u drugoj rečenici.

Pored varijante koja koristi LSA algoritam, Lintean i Rus [11] su razmotrili i pet WordNet metrika nad istim modelom, koristeći POS tagove na već predstavljen način.

TABELA 1: PREGLED ALGORITAMA ZA ODREĐIVANJE SEMANTIČKE SLIČNOSTI KRATKIH TEKSTOVA

Algoritam	Tačnost	Preciznost	Odziv	F-mera	Korišćene sintaksne informacije				Tip
					Redosled reči	POS tagovi	Parsiranje	SRL	
Statistička sličnost									
Mihalcea et al. - LSA	68,40%	69,70%	95,20%	80,50%		+			L
Mihalcea et al. - PMI	69,90%	70,20%	95,20%	81,00%		+			L
Furlan et al. - COALS	70,32%	/	/	/					D
Islam i Inkpen	72,64%	74,70%	89,10%	81,30%	+				D
Lintean i Rus - LSA	73,00%	77,30%	84,00%	80,50%		+			L
Blacoe i Lapata - SDS	73,04%	/	/	82,33%					D
Blacoe i Lapata - DM	73,51%	/	/	82,16%		+	+		T
Topološka sličnost									
Furlan et al. - ConceptNet	68,23%	/	/	/				+	T
Mihalcea et al. - Combined	70,30%	69,60%	97,70%	81,30%		+			L
Li et al.	70,80%	70,30%	97,40%	81,60%			+		T
Oliva et al.	70,87%	74,47%	84,17%	79,02%		+	+	+	T
Fernando i Stevenson	74,10%	75,20%	91,30%	82,40%		+			L
Lintean i Rus - LCH	75,70%	78,30%	87,90%	82,80%		+			L

V. EVALUACIJA I KLASIFIKACIJA

Rezultati svih predstavljenih algoritama, sortirani po tačnosti, prikazani su u tabeli 1. Testiranje algoritama je sprovedeno nad MSRPC korpusom (*Microsoft Research Paraphrase Corpus*) koji sadrži 5801 par rečenica koje su u nekoj meri semantički povezane, ali od kojih su samo neke semantički podudarne tj. parafraze [16]. Ceo korpus je ručno anotiran binarnim ocenama sličnosti koje govore da li posmatrani par rečenica predstavlja parafrazu ili ne.

Glavna mera performansi je tačnost algoritma (*accuracy*), koja se računa kao odnos broja tačno klasifikovanih parova rečenica i ukupnog broja parova u korpusu. Takođe se koriste i sledeće metrike: preciznost (*precision*), koja predstavlja odnos broja tačno prepoznatih parafraza i broja parova rečenica koje je algoritam označio kao parafraze; odziv (*recall*), koji predstavlja odnos broja tačno prepoznatih parafraza i ukupnog broja parafraza u korpusu; i F-mera (*F-measure*) koja se računa kao harmonijska sredina preciznosti i odziva.

Imajući u vidu da se stepen razvijenosti elektronskih jezičkih alata razlikuje od jezika do jezika, kao i da za manje jezike mnogi složeniji alati često nisu dostupni, jasno je da primenjivost algoritama određivanja semantičke sličnosti na određeni jezik umnogome zavisi od sintaksnih informacija koje ti algoritmi koriste. Stoga u tabeli 1 predlažemo novu klasifikaciju ovih algoritama na tri tipa – D, L i T – na osnovu njihove primenjivosti na jezike sa slabije razvijenim jezičkim alatima.

U tip D ulaze algoritmi koji su direktno primenjivi na svaki jezik, jer ne koriste nikakve sintaksne informacije za čije su dobijanje potrebni alati specifični za svaki jezik. Tip L obuhvata algoritme koje je relativno lako moguće primeniti na veći broj jezika, jer od sintaksnih alata koriste samo POS tagere, koji predstavljaju osnovni i najšire rasprostranjeni tip sintaksnih alata. Konačno, tip T sačinjavaju algoritmi koje je teško primeniti na veći broj jezika, jer koriste napredne tehnike sintaksnog procesiranja, kao što su parsiranje i SRL, koje su dostupne u malom broju jezika. Treba napomenuti da je, bez obzira na korišćene sintaksne informacije, kod algoritama topološke sličnosti poseban problem da li u posmatranom jeziku postoje ručno modelovane baze kao što je WordNet, što dodatno ograničava njihovu primenjivost.

Iz tabele 1 se može videti da je broj algoritama D tipa srazmerno mali, i da svi koriste princip statističke sličnosti. Algoritmi L tipa su najrasprostranjeniji. Primetno je da svi noviji algoritmi L tipa postižu podjednake ili bolje tačnosti od algoritama D tipa, što je i očekivano. Međutim, interesantno je da, barem u okviru algoritama topološke sličnosti, algoritmi L tipa ostvaruju bolje rezultate i od algoritama T tipa, koji koriste dublje sintaksne informacije. Jedan od glavnih razloga za ovo leži u činjenici da napredniji sintaksni alati, iako daju dublje sintaksne podatke, takođe proizvode i primetno veći procenat grešaka od jednostavnijih POS tagera.

Od algoritama D tipa najbolje performanse ostvaruje SDS pristup koji su predložili Blacoe i Lapata [12], a veoma blizu mu je i algoritam koji nude Islam i Inkpen [9]. Interesantno je da su Islam i Inkpen [9] ponudili verziju algoritma koja koristi jednostavne sintaksne informacije u vidu broja istih reči koje se javljaju u podudarnim redosledima u oba teksta, ali se pokazuje da dodavanje takve metrike zapravo blago pogoršava rezultate algoritma. Iako je statistički pristup koji predlažu Furlan et al. [10] dosta sličan onome iz [9], oni za treniranje COALS algoritma koriste dosta mali korpus apstrakata članaka sa engleske Vikipedije, te su rezultati očekivano slabiji.

Lintean i Rus [11] nude najbolji algoritam u L klasi, koji je ujedno i najbolji algoritam uopšte, i koji postiže primetno bolju tačnost i veću F-meru od svih ostalih. Topološki algoritmi L tipa generalno ostvaruju više performanse od statističkih. Iako izbor WordNet metrike u topološkim algoritmima zavisi od ostalih elemenata predloženog pristupa, dve metrike se izdvajaju po visokim performansama – JCN, koja najbolje rezultate daje u algoritmima [13] i [15], a drugi najbolji rezultat u [11], i LCH, koja daje state-of-the-art performanse u kombinaciji sa pristupom [11], a drugi najbolji rezultat u [13]. Stariji pristupi, kao što je [8], postižu najbolje rezultate

kombinovanjem svih WordNet metrika sa algoritmima statističke sličnosti, ali je problem ovakvog pristupa što je on dosta spor, jer koristi osam raznorodnih algoritama.

Iako većina algoritama T tipa spada u kategoriju topološke sličnosti, Blacoe i Lapata [12] nude statistički DM algoritam koji ih nadmašuje po performansama. Ipak, čak i takav state-of-the-art algoritam T tipa je tek nešto bolji od daleko jednostavnijeg SDS pristupa iz istog rada, što govori u prilog algoritama D tipa. Gotovo svi algoritmi T tipa potvrđuju da se tačnost sistema može poboljšati davanjem različitih težina sličnostima različitih rečeničnih konstituenata. Algoritam predstavljen od Li et al. [14] postiže najveću tačnost ako se glagolima i imeničkim sintagmama da malo veća težina nego drugim rečeničnim konstituentima. Model koji nude Oliva et al. [15] ostvaruje bolje performanse ako se različitim semantičkim ulogama daju različite težine, pri čemu se predikati ponderuju najvišom ocenom, subjekti i objekti nešto nižom, a priloške odredbe najnižom. Pristup iz [10] ostvaruje najbolje rezultate ako se predikatima da četiri puta veća težina nego subjektima i objektima. Ipak, ovaj pristup postiže niže rezultate od statističkog algoritma iz istog rada, čemu je jedan od uzroka to što mnoge subjekte i objekte čine lična imena koja se ne mogu pronaći u ConceptNet bazi, te se stoga ni ne uzimaju u razmatranje.

VI. Zaključak

Analiza pokazuje da se trenutne state-of-the-art performanse postižu korišćenjem POS tagova u kombinaciji sa mehanizmom topološke sličnosti. Algoritmi koji ne koriste sintaksne informacije žrtvuju određen nivo tačnosti zarad jednostavnosti i široke primenjivosti, dok su algoritmi koji koriste duboke sintaksne informacije još uvek ograničeni tačnošću sintaksnih alata koje koriste.

Jedan od mogućih načina prevazilaženja ovog problema bi bila upotreba jednostavnijih sintaksnih alata uz dublju analizu njihovih rezultata. Naime, predstavljeni algoritmi koriste POS tagove ili kao ulaz za sintaksne alate višeg nivoa, ili u cilju sprečavanja poređenja dve reči koje ne pripadaju istim vrstama reči. Međutim, iako je više radova pokazalo da se tačnost sistema povećava kada se uzmu u obzir prirodne razlike u važnosti različitih rečeničnih konstituenata ([10], [14], [15]), nijedan algoritam nije razmatrao korišćenje tih razlika na POS nivou.

Dodeljivanje različitih težina sličnostima različitih vrsta reči bi uzelo u obzir ovu razliku u važnosti, ali bi se istovremeno, uzevši u obzir visoku tačnost POS tagera, izbegla cena u performansama koja nastaje zbog grešaka kompleksnijih sintaksnih alata. Ovakav pristup, naročito u kombinaciji sa statističkim algoritmima, omogućio bi kreiranje sistema visoke tačnosti koji bi bio lako primenjiv i u jezicima sa ograničenim elektronskim jezičkim alatima.

Literatura

[1] R. Barzilay and K. R. McKeown, "Sentence Fusion for Multidocument News Summarization," *Computational Linguistics*, vol. 31, no. 3, pp. 297–328, Sep. 2005.

[2] S. M. Harabagiu, S. J. Maiorano, and M. A. Paşca, "Open-Domain Textual Question Answering Techniques," *Natural Language Engineering*, vol. 9, no. 3, pp. 231–267, Sep. 2003.

[3] Z. Harris, "Distributional Structure," *Word*, vol. 10, no. 23, pp. 146–162, 1954.

[4] M. P. Marcus, M. A. Marcinkiewicz and B. Santorini, "Building a Large Annotated Corpus of English: The Penn Treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, Jun. 1993.

[5] C. D. Manning, "Part-of-Speech Tagging from 97 % to 100 %: Is It Time for Some Linguistics?" in *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 171–189, 2011.

[6] E. Charniak and M. Johnson, "Coarse-to-fine n-best parsing and MaxEnt discriminative reranking," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 173–180, 2005.

[7] V. Srikumar and D. Roth, "A Joint Model for Extended Semantic Role Labeling," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 129–139, 2011.

[8] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and Knowledge-based Measures of Text Semantic Similarity," in *Proceedings of the 21st National Conference on Artificial Intelligence*, pp. 775–780, 2006.

[9] A. Islam and D. Inkpen, "Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity," *ACM Transactions on Knowledge Discovery from Data*, vol. 2, no. 2, pp. 1–25, Jul. 2008.

[10] B. Furlan, V. Sivački, D. Jovanović, and B. Nikolić, "Comparable Evaluation of Contemporary Corpus-Based and Knowledge-Based Semantic Similarity Measures of Short Texts," *Journal of Information Technology and Applications*, vol. 1, no. 1, pp. 65–72, 2011.

[11] M. Lintean and V. Rus, "Measuring Semantic Similarity in Short Texts through Greedy Pairing and Word Semantics," in *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*, pp. 244–249, 2012.

[12] W. Blacoe and M. Lapata, "A Comparison of Vector-based Representations for Semantic Composition," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 546–556, 2012.

[13] S. Fernando and M. Stevenson, "A Semantic Similarity Approach to Paraphrase Detection," in *Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*, pp. 45–52 , 2008.

[14] L. Li, Y. Zhou, B. Yuan, J. Wang, and X. Hu, "Sentence Similarity Measurement based on Shallow Parsing," in *Proceedings of the Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 487–491, 2009.

[15] J. Oliva, J. I. Serrano, M. D. del Castillo, and Á. Iglesias, "SyMSS: A syntax-based measure for short-text semantic similarity," *Data & Knowledge Engineering*, vol. 70, no. 4, pp. 390–405, Apr. 2011.

[16] B. Dolan, C. Quirk, and C. Brockett, "Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources," in *Proceedings of the 20th International Conference on Computational Linguistics*, 2004.

Abstract

This paper outlines and categorizes ways of using syntax information in a number of algorithms for determining short text semantic similarity. Algorithm performance was evaluated using the results of a paraphrase detection test on the Microsoft Research Paraphrase Corpus. Among the described algorithms and approaches to using syntax information we identify those best suited for application in languages with limited electronic linguistic tools and, with that goal in mind, we propose a new algorithm classification.

EVALUATION AND CLASSIFICATION OF SYNTAX INFORMATION USAGE IN DETERMINING SHORT TEXT SEMANTIC SIMILARITY

Vuk Batanović, Dragan Bojić