Odkrivanje podobnih vsebin s pomočjo pomenskih podpisov in pomenskega stiskanja

Sandi Majninger[1]

[1]*Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko*
Koroška cesta 46, 2000 Maribor
E-pošta: sandi.majninger@um.si

Detecting content similarity using semantic signatures and semantic compression

The biggest issue in plagiarism detection is dealing with obfuscated and translated contents. To support the process of plagiarism detection, we propose a method that compares contents by its semantic meaning. We can achieve that by calculating a semantic signature of each paragraph. For that we use semantic dictionaries. A semantic signature is a set of semantic tags, which represents the appearance of dictionary terms in the content. Since the semantic dictionaries contains the information about hypernyms and meronyms we can also run a semantic compression of calculated signatures. Instead of original terms, we use their hypernyms or meronyms. By that we generalize the meaning of a paragraph. That improves the ability to detect similarity of obfuscated content.

1 Uvod

Pri odkrivanju plagiatov je največji izziv, kako odkriti parafrazirana in prevedena besedila. V pomoč pri odkrivanju takšnih plagiatov, predlagamo postopek za odkrivanje podobnosti vsebin, ki primerja vsebine glede na njihov pomen. To dosežemo z izračunom pomenskih podpisov posameznih odstavkov. Za to lahko uporabimo pomenske slovarje. Pomenski podpis je množica pomenskih značk, ki povedo kateri pojmi iz slovarja se pojavljajo v besedilu. Pomenski slovarji vsebujejo tudi informacije o nadpomenskih pojmih. To lahko izkoristimo za izvajanje pomenskega stiskanja. Namesto osnovnih pojmov uporabimo njihove nadpomenske pojme. S tem posplošimo pomen odstavka in izboljšamo zmožnost zaznave podobnosti v parafraziranih besedilih.

V prispevku bomo najprej na kratko predstavili problem. Nato bomo opisali predlagan postopek za odkrivanje podobnosti v besedilih. Na koncu bomo podali rezultate, ki smo jih dosegli pri prepoznavanju podobnosti med besedili v korpusu PAN [1]. V zaključku bomo izpostavili najpomembnejše ugotovitve in predstavili načrte za nadaljnje delo.

2 Opis problema

Pri odkrivanju plagiatov naletimo na različne vrste plagiatov. Avtorji se poslužujejo povzemanja in parafraziranja besedil ter prevajanja iz tujih jezikov. Da bi odkrili podobnost v takšnih besedilih, potrebujemo način, ki bo znal primerjati besedila glede na njihov pomen. Za uspešno prepoznavo podobnosti potrebujemo najprej način, s katerim bomo opisali pomen besedila. Nato potrebujemo še postopek, ki bo izvedel primerjavo dveh različnih pomenov. Pri tem bomo izvajali klasifikacijo, ki bo za primerjana pomena odločila, ali sta podobna ali ne. Postopek odkrivanja plagiatov je nemogoče popolnoma avtomatizirati, končno presojo zato podajo pristojne osebe oz. inštitucije. Na podlagi tega lahko trdimo, da je pri klasifikaciji priklic pomembnejši od natančnosti, saj bo natančnost naknadno izboljšana pri končni presoji.

3 Odkrivanje podobnosti s pomenskimi podpisi

Podobnost med dokumenti odkrivamo s primerjavo odstavkov. Odstavek v slovnici izraža neko pomensko zaključeno celoto, zato se zdi primeren za primerjavo pomenov besedil. Za vsak odstavek moramo določiti pomenski podpis. To je množica pomenskih značk, ki nam povedo kateri pojmi iz pomenskega slovarja se pojavljajo v odstavku. Nad pomenskimi podpisi lahko izvedemo še pomensko stiskanje, kar pomeni da vse pojme zamenjamo z njihovimi nadpomenskimi pojmi. Informacijo o nadpomenskih pojmih pridobimo iz pomenskih slovarjev. Pri primerjavi dveh odstavkov, primerjamo njuna pomenska podpisa. To smo izvedli tako, da najprej iz obeh pomenskih podpisov izračunamo vektor značilk primerjave. Tega nato klasificiramo v razred podobnih in nepodobnih odstavkov. Za klasifikacijo smo uporabili metodo podpornih vektorjev, ki smo jo učili z nadzorovanim strojnim učenjem.

3.1 Določevanje pomenskih podpisov

Pomenski podpis je množica pomenskih značk. Za določitev pomenskih značk izvedemo pomensko označevanje besedila z uporabo pomenskega slovarja. Uporabili smo pomenski slovar BabelNet [2]. Iz besedila odstavka smo najprej generirali vse besedne n-grame za $n <= 5$. To pomeni, da smo pripravili seznam vseh besed in besednih zvez z največ petimi besedami, ki se pojavljajo v odstavku. Besedne zveze smo normalizirali. Iz njih smo izluščili vse znake, ki niso del angleške abecede in odvečne presledke. Nato smo vse črke spremenili v velike tiskane črke. Na koncu smo še besede znotraj besednih zvez uredili po abecednem vrstnem redu. Na enak način smo normalizirali nazive

pojmov v pomenskem slovarju. Nato smo seznam besed in besednih zvez, ki se pojavljajo v besedilu poiskali v pomenskem slovarju. Za vsako najdeno ujemanje smo tvorili pomensko značko. Ta poleg pojma vsebuje še informacijo o položaju v odstavku. Tega smo izrazili z odmikom od začetka odstavka in dolžino besede oz. besedne zveze. Stisnjene pomenske podpise smo tvorili tako, da smo vse označene pojme zamenjali z njihovimi nadpomenskimi pojmi.

3.2 Izračun vektorja značilk primerjave

Pri primerjavi dveh pomenskih podpisov, iz njiju izračunamo lastnosti, za katere menimo da nam povedo kdaj sta primerjana odstavka podobna. Najbolj osnovna lastnost je število skupnih pojmov v obeh pomenskih podpisih. Izkaže se, da je ta neprimerna za primerjavo odstavkov različnih dolžin. Zato smo raje izračunali faktor ujemanja v obeh odstavkih. To je količnik med številom skupnih pojmov in številom vseh označenih pojmov v odstavku.

Izkaže se, da faktor ujemanja ne zadošča za uspešno razpoznavo podobnosti. Pri primerjavi odstavkov različnih dolžin se pogosto zgodi, da je veliko pojmov iz krajšega odstavka vsebovanih v daljšem odstavku, vendar so razkropljeni in se ne navezujejo med seboj. Zato smo izračunali še gostoto ujemajočega dela. Najprej smo v odstavku določili ujemajoči del besedila. Tega dobimo v okolici povprečne vrednosti odmika skupnih pojmov z oddaljenostjo največ za eno dolžino standardnega odklona odmikov skupnih pojmov. Na tak način zajamemo približno 65% vseh skupnih pojmov. V ujemajočem delu nato izračunamo gostoto kot količnik števila znakov, ki so označeni kot skupnimi pojmi s številom vseh pomensko označenih znakov v ujemajočem delu.

Izračunane lastnosti primerjave zložimo v vektor značilk. Ta ima štiri komponente: faktor ujemanja in gostoto ujemajočega dela za oba primerjana odstavka.

3.3 Klasifikacija vektorjev značilk

Za klasifikacijo vektorjev značilk v razred podobnih in nepodobnih parov odstavkov smo uporabili metodo podpornih vektorjev. Za učenje smo uporabili algoritem zaporedne minimizacije (angl. *sequential minimal optimization*). Gre za metodo strojnega nadzorovanega učenja. Učno množico smo tvorili iz vektorjev značilk, ki smo jih izračunali nad besedili iz učnega dela korpusa PAN [1]. Naključno smo izbrali 600 negativnih in 60 pozitivnih vzorcev. Zaradi majhnega nabora možnih vhodnih vrednosti, takšna velikost učne množice zadošča.

4 Rezultati

Kot rezultat smo merili uspešnost klasifikatorja za prepoznavo podobnosti v testnem delu korpusa PAN [1]. Ker učno množico izbiramo naključno, smo postopek učenja in razpoznavanja ponovili desetkrat, navedli pa povprečne rezultate. Ker je za področje končne uporabe priklic pomembnejši od natančnosti, smo navedli zgolj število odkritih in neodkritih podobnosti. Da smo povečali stopnjo zaznanih podobnosti, pri učenju metode podpornih vektorjev nismo uporabili pomenskega stiskanja. Nato smo pri prepoznavanju s pomenskim stiskanjem povečali faktorje ujemanja in s tem povečali število odkritih podobnosti, pri čemer smo žrtvovali natančnost postopka. V tabeli 1 so prikazani rezultati meritev na različnih tipih plagiatov iz korpusa PAN [1].

Tabela 1: Rezultati meritev uspešnosti postopka

Vrsta plagiata	Št. odkritih	Št. neodkritih
Prepisovanje	1204	2
Parafraziranje	1241,2	50,8
Ciklično prevajanje	1259,9	48,1
Povzemanje	234	2
Skupaj	**3939,1**	**102,9**

5 Sklep

Iz rezultatov v tabeli 1 je razvidno, da smo uspeli doseči visoko stopnjo priklica. Visoka je tudi stopnja klasifikacijske točnosti. Zaradi večje količine napačno zaznanih podobnosti, je natančnost nizka. Kljub temu je postopek primeren za uporabo kot pripomoček pri odkrivanju plagiatov v novo nastalih delih. Metoda podpornih vektorjev nam omogoča, da izračunamo verjetnost pripadnosti razredu. Na podlagi tega bi lahko določili stopnjo podobnosti in tako vrnili le dokumente z najvišjo stopnjo podobnosti.

Kot možnosti za izboljšave predlagamo povečanje stopnje natančnosti. To bi lahko dosegli s pametnejšo izbiro nadpomenskih pojmov pri pomenskem stiskanju.

Zahvala

Zahvaljujem se mentorju doc. dr. Milanu Ojsteršku za nasvete pri raziskovalnemu delu. Hvala tudi sodelavcu mag. Janezu Brezovniku za začetne ideje.

Literatura

[1] Potthast M., Hagen M., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., Stein, B. Overview of the 5th International Competition on Plagiarism Detection, CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, Spain (2013)

[2] Navgili, R., Ponzetto, S. P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193, (2012), str. 217-250