Ekstrakcija ključnih besed filmov iz podnapisov

Jan Popič, Timi Ornik, Nejc Planer, Borko Bošković, Janez Brest

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko,
Koroška cesta 46, 2000 Maribor
E-pošta: jan.popic1@um.si

Keyword extraction from movie subtitles

With large amounts of video content it is important to have as much metadata as possible. Important metadata in video content includes keywords which summarize the essential elements in a given video. In this paper, we present a framework for an automated approach to keyword extraction from video content using their associated subtitles. Framework is composed of three primary building blocks: preprocessing, named-entity recognition and keyword extraction. The goal of the first two blocks is the removal of information irrelevant to the keywords. In the final stage, the actual keywords are extracted from the preprocessed subtitles. In the experiment, we compare the keywords extracted with our framework to a set of predefined keywords obtained from the Internet Movie Database website. We test our method on Slovene language and compare it to a similar English model.

1 Uvod

V današnjih dneh je na voljo ogromna količina video vsebin, le-te pa postajajo del našega vsakdana. Ker pa je v tej poplavi vsebin pogosto težko najti takšno, ki bi nas zanimala, je posebej pomembno, da so dobro označene z metapodatki. Eden izmed ključnih metapodatkov o video vsebini so ključne besede. Z avtomatsko ekstrakcijo ključnih besed lahko neizmerno pohitrimo sicer ročno označevanje. Ključne besede so uporabne predvsem za priporočilne sisteme, pripise in oznake vsebin in povezovanje s socialnimi omrežji.

Ekstrakcija ključnih besed je definirana kot postopek, ki vhodnemu dokumentu dodeli nekaj besed ali besednih zvez, s katerimi najbolje opišemo njegovo vsebino [1]. Ker ima večina video vsebin v današnjih dneh vsaj neko obliko pripadajočih podnapisov, smo se v našem delu osredotočili samo na ekstrakcijo ključnih besed iz slovenskih podnapisov. Zaradi najlažjega dostopa do podnapisov in obstoječih metapodatkov o vsebini, se bomo omejili samo na filme. Enak pristop bi lahko preslikali na vsako video vsebino, ki ima podnapise.

2 Sorodna dela

V [2] so se avtorji ukvarjali z ekstrakcijo ključnih besed iz prepisov (transkripcije) sestankov. Avtorji so izpostavili nekaj pomembnih razlik med prepisi govora in navadnim pisanim besedilom (knjige, članki, prispevki, itd.), ki veljajo tudi za našo domeno analize podnapisov video vsebin:

1. V nenačrtovanem pogovoru sta v stavku povprečno dve pomenski besedi, za razliko od pisanega besedila, kjer je v posameznem stavku od 4 do 6 pomenskih besed. To lahko predstavlja problem osnovnim algoritmom ekstrakcije ključnih besed, ki temeljijo na frekvenci pojavitev.
2. V govorjenem besedilu je pogovor manj strukturiran. Manjkajo tudi dodatne informacije, ki jih lahko zasledimo v drugih pisanih besedilih (naslovi, odstavki, poglavja, itd.).
3. Za razliko od pisanih besedil, kjer je navadno en sam avtor, imajo v prepisih različni govorci različne sloge govorjenja in različno rabo besed.
4. Ker je struktura govora ohlapna in ni zagotovila, da je pravilna, se lahko pojavijo težave z označevanjem besednih vrst (angl. *POS tagging* oz. *part-of-speech tagging*).

Avtorji prispevka [3] so po našem znanju edini, ki so uporabili dodatne informacije na voljo v podnapisih. Njihov pristop je deljen na 5 osnovnih gradnikov: predobdelava, prepoznava imenskih entitet, ločevanje pogovorov, ekstrakcija ključnih besed in prilagoditev uteži ključnih besed. Poročali so o 11,2 % izboljšavi metrike *F1* z algoritmom TextRank [4] in uporabo prej navedenih korakov v primerjavi z uporabo algoritma TextRank nad celotno vsebino podnapisov.

3 Zasnova algoritma

Naš primarni cilj je bil izdelava sistema za ekstrakcijo ključnih besed filmov iz podnapisov, ki so v slovenskem jeziku. Za enostavnejše nadaljnje delo in preprosto ter intuitivno uporabo smo želeli izdelati ogrodje, ki omogoča enostavno konfiguracijo komponent (odstranjevanje in spremembo posameznih delov ogrodja).

Pri snovanju našega pristopa smo se zgledovali po podobnem pristopu za angleški jezik, ki so ga predstavili avtorji v [3]. Vendar naš pristop ne vsebuje ločevanja pogovorov, prav tako se v koraku prepoznave imenskih entitet nismo omejili samo na unikatne besede, ki se nikoli ne pojavijo z malo začetnico, temveč smo dovolili malo

število ponovitev zaradi morebitnih slovničnih in besednih napak v samih podnapisih. Naš pristop je razdeljen na tri osnovne gradnike:

a) Predobdelava

V tej fazi smo iz podnapisov izločili nepotrebne informacije: vse značke HTML (barvne, glasbene, način govora, označbe dogajanja), znake, ki nakazujejo spremembo govorca ('-') in dvojne presledke. Prav tako smo iz vsake datoteke s podnapisa odstranili zadnja dva podnapisa, ker le-ta ponavadi vsebujeta podatke o avtorju prevoda, kar ni pomembno za ekstrakcijo ključnih besed. Nato smo odstranili nepomembne in nepotrebne besede s pomočjo v naprej definiranega NLTK (angl. *Natural Language Toolkit*) seznama nepomembnih besed [5].

b) Prepoznava imenskih entitet (PIE)

V tej fazi smo filtrirali besedne vrste ter ločili imenske entitete (predmete iz resničnega sveta, na primer osebe ali lokacije, ki jih lahko označimo z lastnimi imeni), ki so primerne za ključne besede od tistih, ki to niso.

Najprej smo nad vsebino posameznega podnapisa zagnali algoritem za označevanje besednih vrst, da smo pridobili informacije o strukturi stavka. Iz posameznega stavka se nato izbrišejo vse besede, ki ne predstavljajo želenih besednih vrst (želene besedne vrste so uporabniško nastavljiv parameter).

Izmed besed, ki so ostale v stavku, se kot kandidatke označijo vse besede, ki predstavljajo lastna imena in so se z malo začetnico ponovile največ dvakrat (na ta način omilimo morebitne sintaktične napake v podnapisih).

Za vse kandidatke smo s pomočjo semantičnega leksikona slovenščine sloWNet [6] preverili, ali je njihova leksikalna domena na seznamu dovoljenih (dovoljene domene so uporabniško nastavljiv parameter). Vse, ki niso ustrezale seznamu dovoljenih leksikalnih domen, smo odstranili iz stavkov.

c) Ekstrakcija ključnih besed

Za samo ekstrakcijo ključnih besed smo uporabili dva različna algoritma (TD-IDF [7] in TextRank), ki smo ju nato med seboj primerjali po uspešnosti.

3.1 Zbiranje podatkov

Korak prepoznave imenskih entitet potrebuje nekaj modelov in korpusov (nabor besedil) za pravilno deljenje besed (angl. *tokenizer*), označevanje besednih vrst in leksikalno analizo. Uporabili smo sledeče modele in korpuse:

- seznam nepomembnih slovenskih besed iz NLTK,
- jezikovni model Punkt za deljenje povedi iz NLTK,
- korpus "MULTEXT-East 1984" za označevanje besednih vrst iz NLTK,
- semantični leksikon Open Multilingual Wordnet za dodatne informacije in povezavo s sloWNet iz NLTK,
- model Word2vec "Slovenian CoNLL17 corpus" iz NLPL [8] za mero podobnosti in
- semantični leksikon slovenščine za pridobitev domen in sinonimov sloWNet.

Za uspešno analizo rezultatov potrebujemo dovolj veliko testno množico podnapisov in ključnih besed, ki predstavljajo nabor ključnih besed, ki so jih določili uporabniki in ocenjevali za najbolj primerne. Nabora ključnih besed za filme nismo našli v slovenskem jeziku, zato smo ga pridobili v angleškem jeziku iz spletne strani IMDb [9] (urejene po ocenah uporabnikov). Ključne besede smo nato prevedli v slovenski jezik s pomočjo prevajalnika Microsoft Translator [10]. Zaradi prevajanja smo morali vse ključne besede spremeniti v male črke. Podnapise smo pridobili iz portala Podnapisi.NET [11], kjer pa ni nujno, da obstaja podnapis za vsak izbran film, ta pa je lahko tudi napačen.

Zaradi samega prevajanja in načina pridobivanja podnapisov smo posledično uvedli nekaj napak v naš korpus. Dodatne napake se lahko pojavijo tudi v samih podnapisih, navadno tistih, ki ne izhajajo iz uradnih studiov. Izbrali smo 100 najbolje ocenjenih filmov iz IMDb in pridobili en podnapis ter 60 prevedenih ključnih besed za vsak film. Primer nekaj ključnih besed je viden v tabeli 1.

Tabela 1: Primeri ključnih besed iz IMDb

Film	Prevedene ključne besede
The Matrix (1999)	simulirano resničnost, umetna realnost, prerokba, programer, po apokalipsi, borilnih veščin, hacker, distopija, človeštvo v nevarnosti, tabletke
Fight Club (1999)	presenečenje konča, boj, proti obratu, multiple osebnostne motnje, nespečnost, proti skladnosti, na podlagi novih, skupinsko zdravljenje, proti kapitalizmu, pretep

4 Eksperiment

Naš eksperiment smo zasnovali tako, da smo za vsak film iz podnapisov izluščili ključne besede z našim algoritmom, te pa smo nato primerjali s tistimi pridobljenimi iz IMDb. Za pravilno pridobljeno ključno besedo smo upoštevali dobesedno ujemanje, sinonime pridobljene iz sloWNet in mero podobnosti. Pri tem nismo preverjali ustreznih velikih začetnic, saj ima naša zbirka pridobljenih ključnih besed izključno male črke. Za ovrednotenje smo naš pristop smatrali kot klasifikator v en razred. Iz matrike klasifikacije (angl. *confusion matrix*) smo nato izračunali metriko $F1$, natančnost P (angl. *precision*) in priklic R (angl. *recall*).

Analizirali smo kombinacije algoritmov za ekstrakcijo besed (TF-IDF in TextRank) z in brez modula prepoznave imenskih entitet PIE (glej poglavje 3), saj nas je zanimal vpliv tega modula na kvaliteto ključnih besed.

V modulu PIE smo v tem primeru obdržali samo lastna imena, samostalnike, pridevnike in števila. Iz lastnih imen smo v nadaljevanju izluščili samo tiste, ki imajo veliko začetnico in se ne pojavijo več kot dvakrat z malo začetnico. Izluščene besede smo preverili s sloWNet in odstranili vse, ki ne spadajo v izbrano domeno (obdržane domene *geography, chemistry, telecommunication, astronomy, politics, animals* so bile izbrane s subjektivno oceno primernosti ključnih besed). Pri odstranjevanju neželenih

besednih vrst se v povprečju odstrani 69 % unikatnih besed, pri odstranjevanju neželenih leksikalnih domen - z našimi obdržanimi domenami - pa 1,5 %. Leksikalne domene imajo vpliv predvsem na tip pridobljenih ključnih besed (subjektivna definicija ključne besede), ne toliko na kvaliteto. Skupno se v povprečju odstrani 71 % vseh besed (vključno s ponovitvami).

Tabela 2: Primerjava različnih kombinacij algoritmov v %

Metoda	*F1*	*P*	*R*
TF-IDF	1,468	1,468	1,469
TF-IDF + PIE	**2,320**	**2,319**	**2,321**
TF-IDF + PIE + S	8,061	8,061	8,061
TextRank	1,456	1,461	1,452
TextRank + PIE	1,979	1,988	1,969
TextRank + PIE + S	**8,770**	**8,760**	**8,780**

PIE - Prepoznava imenskih entitet, S - Sinonimi

Vse kombinacije algoritmov in njihovi rezultati so vidni v tabeli 2.
Največjo natančnost brez upoštevanja sinonimov je dosegel algoritem TF-IDF. Brez modula PIE ima algoritem TF-IDF le majhno prednost pred algoritmom TextRank, če vključimo modul PIE se oba rezultata izboljšata, razlika med rezultatoma obeh algoritmov pa postane bolj očitna. Z upoštevanjem sinonimov pa ima TextRank še večjo natančnost.

Nadalje nas je zanimal vpliv nabora besednih vrst, ki se v modulu PIE izbrišejo. Za analizo tega smo izbrali algoritem TextRank, saj se je izkazalo, da dosega najboljše rezultate za naš problem, filtriranje domen pa smo izključili. Kot je razvidno iz tabele 3, imajo največji vpliv na rezultat glagoli, saj se ti v podnapisih velikokrat pojavijo, so pa le redko ključne besede.

Tabela 3: Vpliv nabora besednih vrst v %

Obdržani členi	*F1*	*P*	*R*
Vsi + S	5,435	5,429	5,441
NN, A, NUM, V + S	7,066	6,959	7,176
NN, A, NUM + S	8,660	8,647	8,674
NN, A + S	**8,770**	**8,760**	**8,780**

NN - Samostalniki, A - Pridevniki, NUM - Števila, V - Glagoli, S - Sinonimi

Na koncu nas je zanimala še primerjava z upoštevanjem podobnosti izluščenih ključnih besed s pridobljenimi na najboljših kombinacijah iz prejšnjih eksperimentov (algoritem TextRank, modul PIE s samostalniki in pridevniki ter upoštevanje sinonimov).

Podobnosti besed v vmesnih eksperimentih nismo upoštevali s ciljem večje natančnosti izbire in v izogib popačenja vmesnih rezultatov. Prav tako, bi v eksperimentu, kjer smo primerjali besedne vrste, vplival na izbiro besednih vrst, v primeru glagola bi bili rezultati nepravilno višji, saj si je veliko glagolov in samostalnikov zelo podobnih, vendar se glagoli običajno ne uporabljajo kot ključne besede.

Podobnost besed predstavimo s pragom podobnosti pridobljenim s pomočjo modela Word2vec. S tem želimo zgladiti napake pri prevajanju in poskušamo upoštevati subjektivnost in abstraktnost ključnih besed. Ta korak je uporabljen izključno za grobo analizo rezultatov in ne vpliva na delovanje algoritma. Prag z vrednostjo 0,7 v tabeli 4 predstavlja vrednost mere podobnosti, nad katero se ključna beseda šteje kot pravilni zadetek. Prag 0,7 je bil določen s preliminarni poizkusi in predstavlja smiselno vrednost pri kateri se pomensko podobne besede označijo kot sprejete, pomensko različne besede pa se še vedno zavrnejo. Nižji prag bi sicer prinesel boljše kvantitativne rezultate, a same ključne besede ne bi bile smiselne.

Tabela 4: Rezultati najboljše kombinacije v %

Podobnost	*F1*	*P*	*R*
1,0	8,770	8,760	8,780
0,7	**13,869**	**13,870**	**13,869**

4.1 Opisna analiza

Tabela 5: Primeri izluščenih ključnih besed

Film	Izluščene besede
The Matrix (1999)	vrata, **prerokba**, morfej, konec, časa, poveljnik, **programer**, **program**, **človek**, matrico, sion
Fight Club (1999)	gospod, tyler, življenje, stanovanje, **rak**, tylerja, daj, singer, ime, **pretep**

Primere izluščenih ključnih besed vidimo v tabeli 5. S krepko so označene besede, ki se v pomenu ujemajo s ključnimi besedami iz IMDb. Pojavljajo se imena ključnih oseb v filmu, tudi mesta in prostori ter ostale pogoste besede. Vidimo, da so v prevedenih ključnih besedah iz zbirke IMDb v tabeli 1, veliko bolj opisne besede, ki se v samih podnapisih nikoli ne pojavijo in jih je z našim pristopom nemogoče pridobiti. Te opisujejo dele filma in prizore, npr. presenetljiv konec in grafično nasilje, ali pa opisujejo druge podatke o filmu kot npr. “trilogija” ali “ki temelji na romanu”.

Kljub temu, je naše subjektivno mnenje, da z našim pristopom pridobljene besede relativno dobro opisujejo filme, čeprav je v primerjavi z IMDb majhno število zadetkov.

4.2 Primerjava

Avtorji prispevka [3] so dosegli najboljše rezultate z algoritmom TextRank, prepoznavo imenskih entitet in analizo pogovorov.

Zaradi različnega jezika in korpusov ne moremo neposredno primerjati rezultatov. Vsaka metoda je prilagojena svojemu jeziku a sledi istim načelom ekstrakcije ključnih besed. Primerjavo zglajenih rezultatov lahko vidimo v tabeli 6.

Tabela 6: Primerjava rezultatov v %

Metoda	*F1*	*P*	*R*
Prispevek [3] *(angleščina)*	16,99	30,30	11,80
Naš pristop *(slovenščina)*	13,869	13,870	13,869

4.3 Diskusija

Predlagan pristop k ekstrakciji ključnih besed ima dve veliki omejitvi:

a) Subjektivnost

Ključne besede za določen film so zelo subjektivne narave. Ni standardne definicije kaj je in kaj ni ključna beseda, saj je to odvisno od vsakega posameznika. To onemogoči kvantitativno ocenitev kvalitete algoritmov. Iz tega razloga sta naša eksperimenta uporabna samo za ocenitev vpliva posameznih delov in ne za ocenitev kvalitete pridobljenih ključnih besed.

b) Abstraktnost

Ključne besede pogosto vsebujejo abstraktne pojme, ki opisujejo film (npr. "razmerje oče-sin"). Te se skoraj nikoli ne pojavijo znotraj govora v filmu, kar onemogoči ekstrakcijo takšnih besed z našim pristopom.

5 Zaključek

V prispevku smo predstavili ogrodje za avtomatizirano ekstrakcijo ključnih besed filmov iz njihovih podnapisov. V našem pristopu v prvi fazi odstranimo nepotrebne značke HTML, ki jih lahko zasledimo v podnapisih, in besede brez pomena. Nato sledi faza filtriranja, kjer analiziramo in označimo besedne vrste. Posamezne stavke filtriramo tako, da odstranimo besedne vrste za katere ocenimo, da ne predstavljajo dobrih ključnih besed. Sledi analiza imenskih entitet, pri kateri odstranimo vsa lastna imena, ki ne ustrezajo domenam za katere ocenimo, da so nepomembne za ključne besede. V zadnji fazi ekstrakcije ključnih besed z uveljavljenimi algoritmi iz filtriranih podnapisov izluščimo ključne besede filma.

Za analizo kvalitete in konfiguracije (izbira besednih vrst za odstranitev, domen imenskih entitet za odstranitev in izbira algoritma) našega ogrodja smo pridobili slovenske podnapise filmov iz Podnapisi.NET in pred-definirane angleške ključne besede posameznih filmov iz baze IMDb. Ključne besede smo prevedli v slovenski jezik. Problem ekstrakcije ključnih besed smo, za potrebe vrednotenja, obravnavali kot klasifikacijski problem. Ključne besede iz zbirke IMDb in njihove sinonime smo obravnavali kot pravilne.

Analizirali smo kvaliteto izluščenih ključnih besed z algoritmoma TF-IDF in TextRank, pri čemer smo dodatno analizirali vpliv faze filtriranja (PIE). Rezultati nakazujejo, da dosega najboljše rezultate kombinacija TextRank + PIE z upoštevanjem sinonimov.

Dodatno smo analizirali tudi vpliv nabora besednih vrst, ki se odstranjujejo v fazi PIE. Iz rezultatov je razvidno, da daje naš pristop najboljše rezultate, če se odstranijo vse besedne vrste razen lastnih imen (filtriranje domen), samostalnikov in pridevnikov.

V povprečju smo dosegli ujemanje 13,96 ključnih besed na film, v najslabšem primeru 3 besede, v najboljšem primeru 33 besed.

5.1 Nadaljnje delo

Za ovrednotenje kvalitete pridobljenih ključnih besed s predlaganim pristopom bi potrebovali subjektivne ocene. Te bi lahko pridobili z izvedbo vprašalnika nad določeno populacijo.

Predvidevamo, da bi dosegli bolj smiselno razvrščene ključne besede z vpeljavo dodatne faze analize individualnih pogovorov. V tej fazi bi razdelili podnapise na individualne pogovore, te pa ovrednotili glede na pomembnost. Ključne besede, pridobljene iz pomembnih pogovorov, bi imele večjo težo kot tiste pridobljene iz nepomembnih pogovorov.

Morebitno izboljšanje bi lahko dosegli tudi z izbiro drugih algoritmov za ekstrakcijo ključnih besed in podrobnejšo analizo vplivov naših parametrov (množica dovoljenih leksikalnih domen, največ dovoljenih ponovitev, ...).

Zahvala

J. Brest in B. Bošković priznavata financiranje prispevka s strani Javne agencije za raziskovalno dejavnost Republike Slovenije, raziskovalni program P2-0041 – Računalniški sistemi, metodologije in inteligentne storitve.

Literatura

[1] Slobodan Beliga. Keyword extraction: a review of methods and approaches. *University of Rijeka, Department of Informatics, Rijeka*, pages 1–9, 2014.

[2] Fei Liu, Feifan Liu, and Yang Liu. A supervised framework for keyword extraction from meeting transcripts. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19:538 – 548, 04 2011.

[3] Matúš Košút and Marián Šimko. Improving keyword extraction from movie subtitles by utilizing temporal properties. In Rūsiņš Mārtiņš Freivalds, Gregor Engels, and Barbara Catania, editors, *SOFSEM 2016: Theory and Practice of Computer Science*, pages 544–555, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg.

[4] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[5] NLTK. NLTK Data. Dostopno na `https://github.com/nltk/nltk_data`, 2019. [Dostopano 5. maja 2020].

[6] Darja Fišer. Semantic lexicon of slovene sloWNet 3.1, 2015. Slovenian language resource repository CLARIN.SI.

[7] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[8] Nordic Language Processing Laboratory. NLPL word embeddings. Dostopno na `http://vectors.nlpl.eu/repository/`. [Dostopano 23. junija 2020].

[9] IMDb. `https://www.imdb.com/`. [Dostopano 18. junija 2020].

[10] Microsoft Translator. `https://www.microsoft.com/en-us/translator/`. [Dostopano 18. junija 2020].

[11] Podnapisi.NET. `https://www.podnapisi.net/`. [Dostopano 8. maja 2020].