Ivan Dunđer
Sanja Seljan
Hrvoje Stančić
Filozofski fakultet u Zagrebu

KONCEPT AUTOMATSKE KLASIFIKACIJE REGISTRATURNOGA I ARHIVSKOGA GRADIVA

UDK 004.9:930.251

Pregledni znanstveni rad

Sustavi za upravljanje dokumentima i zapisima (EDRMS), koji su najčešće dijelovi sveobuhvatnijeg sustava za upravljanje korporacijskim sadržajima (ECMS) zahvaćaju dokumente i zapise koji izvorno nastaju u digitalnom obliku kao i one koji su digitalizirani. Dok je izvorno digitalne zapise relativno jednostavno opisati tijekom njihova nastanka te im pridodati sve potrebne metapodatke, do problema dolazi kod onih koji u sustav ulaze prolazeći postupak digitalizacije. Ukoliko je riječ o velikoj količini gradiva, pri čemu su dokumenti raznorodni i nemaju neka jedinstvena ili ponavljajuća obilježja, tada nije jednostavno odrediti o kojem je dokumentu riječ, ispravno ga klasificirati i pridodati mu metapodatke. Autori analiziraju i prikazuju mogućnosti rješenja koja pripadaju području statistički utemeljenih jezičnih tehnologijai istražuju njihovu moguću primjenu u području (polu)automatske klasifikacije registraturnoga i arhivskoga gradiva. U radu su objašnjena osnovna polazišta pojedinih metoda, mogućnosti automatske ekstrakcije teksta, metode statističke obrade te postavljanje osnove za (polu)automatsku klasifikaciju. Autori prikazuju rezultate testiranja primijenjenih metoda na konkretnome arhivskom gradivu i zaključuju o mogućim budućim pravcima istraživanja.

Ključne riječi: *automatska klasifikacija, računalna obrada prirodnoga jezika, statističke metode, digitalizacija, arhivsko gradivo*

Uvod

Suvremeno poslovanje u digitalnoj okolini nameće potrebu za uvođenjem i korištenjem sustavâ za elektroničko upravljanje sadržajima (engl. *Electronic Content Management System* – ECMS, ponekad i *Digital Asset Management* – DAM sustavi), elektroničko upravljanje dokumentima (engl. *Electronic Document Management System* – EDMS), elektroničko upravljanje zapisima, odnosno elektroničkih spisovodstvenih sustava (engl. *Electronic Records Management System* – ERMS) koji su najčešće dijelovi cjelovitog sustava na razini ustanove koji upravljaju ukupnošću tzv. korporativnih sadržaja (engl. *Electronic Corporate Management System* – ECMS[1]). Taj sustav u cjelini ili njegovi navedeni podsustavi, koji se u praksi pojavljuju i kao samostalni sustavi, upravljaju digitalnim dokumentima bez obzira na to jesu li oni izvorno nastali u digitalnome obliku ili su digitalizirani. Ipak, u kontekstu njihove obrade, izvorno digitalni dokumenti i zapisi jednostavniji su za upotrebu u smislu njihova jednostavnijeg opisa metapodatcima te pretraživosti cijelog teksta. Problemi se javljaju s dokumentima i zapisima koji su izvorno nastali u analognom obliku ili u digitalnom obliku pa su bili otisnuti te je potrebna njihova ponovna digitalizacija. U nastavku ovoga rada fokus će biti upravo na toj kategoriji i problemima automatske identifikacije dokumenta i njegova klasificiranja.

Automatska identifikacija digitaliziranih dokumenata

U praksi se razlikuju najmanje dva tipa dokumenata – strukturirani i nestrukturirani tip. Uz ta dva tipa još se mogu razlikovati i polustrukturirani dokumenti. Strukturirani dokumenti su oni dokumenti kod kojih je poznata struktura, tj. pojedini elementi dokumenta uvijek se nalaze na istom mjestu. Primjer takvog tipa dokumenta je uplatnica kod koje je uvijek poznata, primjerice, pozicija imena i prezimena uplatitelja. Tijekom postupka digitalizacije strukturiranih tipova dokumenta moguće je odrediti poziciju s koje se pojedini podatak treba iščitati. Ta se pozicija određuje pozicijom u zamišljenom XY koordinatnom sustavu koji čini prostor dvodimenzionalnoga dokumenta. S druge strane, polustrukturirani dokumenti su oni dokumenti kod kojih je poznato da se informacije važne za njihovu

[1] Skraćenica ECMS, ovisno o kontekstu, može imati dva različita značenja, pa tako „C“ može značiti ‘Content’ ili ‘Corporate’.

ispravnu identifikaciju nalaze u naslovima ili podnaslovima dokumenta iako pritom nije poznata njihova točna pozicija. S obzirom na to da su, u pravilu, naslovi i podnaslovi grafički uočljiviji, npr. otisnuti većim fontom i/ili masnim slovima, jednostavno je pronaći takve elemente u digitaliziranome dokumentu i iz njih iščitati informaciju te na temelju nje identificirati dokument. Konačno, nestrukturirani dokumenti su oni dokumenti kod kojih struktura nije unaprijed poznata. Kod njih je u pravilu slučaj da su dokumenti koji pripadaju istoj razredbenoj skupini s grafičkog stajališta vrlo različiti, pri čemu niti opseg dokumenata nije jednak. Stoga se kod njih ne mogu primijeniti dvije do sada objašnjene tehnike kojesu pogodne za identifikaciju strukturiranih i polustrukturiranih dokumenata. Uz to valja napomenuti da je Merrill Lynch, odjel Bank of America, 1998. procijenio da se 80 – 90% svih poslovnih informacija nalazi u nestrukturiranome obliku, dok ComputerWorld procjenjuje da ih je u poslovnim organizacijama 70 – 80%[2]. Drugim riječima, najveći dio poslovne dokumentacije, a samim time kasnije i arhivskoga gradiva, nalazi se upravo u nestrukturiranome obliku, tj. nemaju zajednička identifikacijska svojstva po kojima bi ih se lako moglo prepoznati, pa je potreba za njegovom automatskom identifikacijom i klasifikacijom tim veća.

Tablica 1. Karakteristike i primjeri dvaju osnovnih tipova dokumenata

Tip dokumenta	**Strukturirani**	**Nestrukturirani**
Karakteristike	Podatci za identifikaciju i klasifikaciju dokumenta uvijek se nalaze na istim mjestima	Podatci za identifikaciju i klasifikaciju dokumenta nalaze se bilo gdje u dokumentu
Primjeri	Uplatnica Zahtjev za isplatu štete po osiguranju Obrazac za prijavu – potreba za radnikom	E-poruka ili pismo Dokument, pregledni dokument (engl. *whitepaper*), dopis Izvještaj

[2] Quigley, R. Big data in commodity markets. Business Opportunity or Another Fad? *Commodities Now*,ožujak 2014, str. 61–65. URL:http://www.datagenicgroup.com/wp-content/uploads/2014/06/Big_Data_In_Commodity_Markets.pdf (3.5.2015.)

Poznata komercijalna rješenja kojima se kvalitetno mogu identificirati i klasificirati strukturirani i polustrukturirani dokumenti su *Abbyy FlexiCapture*, *EMC Captiva Capture*, *Ephesoft Smart Capture*, *IBM Datacap*, *Kodak Capture*, *Kofax*, *OnBase AnyDoc* i drugi. Ona mogu kvalitetno klasificirati dokumente prema:

- *grafičkom izgledu* – dokument se analizira vizualno prema grafičkim elementima koji moraju biti prisutni, te njihovoj veličini i poziciji;
- *iščitavanjem sadržaja s unaprijed određene pozicije* (zonskim OCR-om, engl. *Optical Character Recognition*; ICR-om, engl. *Intelligent Character Recognition*; OMR-om, *Optical Mark Recognition*, tj. optičkim prepoznavanjem tiskanih i pisanih znakova te oznaka), najčešće kod strukturiranih dokumenata – određuje se geometrijska pozicija u dvodimenzionalnom prostoru dokumenta te se s te, uvijek jednake, pozicije iščitava relevantna informacija (npr. urudžbeni broj i klasa koji se uvijek nalaze npr. u desnom gornjem kutu dokumenta na nekoj unaprijed definiranoj poziciji u XY koordinatnom sustavu koji reprezentira dokument);
- *ključnim riječima* iz jasne strukture dokumenta, najčešće kod polustrukturiranih dokumenata – prepoznaje se struktura dokumenta u smislu naslova i podnaslova te se na temelju riječi koje se tamo pojavljuju, a ujedno su prisutne u nekom, unaprijed određenom referentnom popisu, klasificira dokument;
- *fizičkoj veličini skenirane stranice* – detekcijom fizičke veličine dokumenta može se odrediti vrsta dokumenta samo ako su svi dokumenti koji se trebaju klasificirati jedinstvene veličine (npr. uplatnice se veličinom razlikuju od kupoprodajnih ugovora); tomse metodom često koristi u kombinaciji s iščitavanjem sadržaja s unaprijed određene pozicije, jer ako se detekcijom veličine dokumenta može odrediti da je riječ o uplatnici, onda je moguće s točno određene pozicije iščitati podatke koji su potrebni za ispravnu klasifikaciju);
- *prepoznavanjem crtičnih kodova (barkodova) i separatora* – kod složenih, odnosno nestrukturiranih dokumenata koje nije moguće automatski prepoznati nekom od prethodnih metoda potrebno je da djelatnik prepozna o kojoj je vrsti dokumenta riječ te na njega postavi jednoznačni element. Uobičajeno je riječ o dodavanju naljepnice s crtičnim kodom na točno određenu poziciju ili ubacivanju zasebnih stranica s informacijama prije

dokumenta (tzv. separatori) koje će skener prilikom skeniranja automatski prepoznati i dokument koji slijedi nakon separatora ispravno klasificirati. Tu je potreban znatan vremenski angažman djelatnika.

Kada se dokumenti jednom ispravno identificiraju, tada se, s unaprijed definiranih pozicija za svaki poznati tip dokumenta, iščitavaju informacije na temelju kojih će se ti dokumenti automatski ili poluautomatski, tj. uz potvrdu djelatnika-operatera, smjestiti u ispravne razredbene skupine, tj. klasificirati. Kao što je ranije naznačeno, problem predstavljaju nestrukturirani dokumenti. Stoga se u nastavku opisuju pristupiekstrakciji termina i kolokacija, kojima se moguće koristiti u (polu)automatskoj identifikaciji i klasifikaciji dokumenata, indeksiranju, u izradi jednojezičnih i dvojezičnih rječnika, u sustavima za pretraživanje informacija i strojnom prevođenju,[3] generiranju teksta[4] i dr., koji bi mogli biti primijenjenina nestrukturiranim dokumentima.

Pristupi identifikaciji nestrukturiranih dokumenata

U procesu ekstrakcije informacija moguće je koristiti se pristupima koji se temelje na statističkim, jezičnim ili hibridnim modelima. Ideja korištenja takvih modela u automatskoj identifikaciji nestrukturiranih dokumenata omogućila bi pronalaženje i jezičnu obradu informacijate određivanje klasifikacijske oznake. Da bi se tako nešto postiglo, sustav se mora „istrenirati" kako bise prepoznao sadržaj nestrukturiranoga dokumenta i odredila njegova klasifikacijska oznaka.

Mnoga tijela državne i javne uprave, organizacije i kompanije provode digitalizaciju svoje poslovne dokumentacije. Istovremeno, tijekom redovitoga poslovanja, nastaje i nova poslovna dokumentacija u papirnatome i digitalnome obliku. Svaki dokument uglavnom je različitoga grafičkog izgleda i strukture, pa se može zaključiti da je dominantno riječ o nestrukturiranim dokumentima.

Na ovome mjestu potrebno je naglasiti da se potreba za identifikacijom nestrukturiranih dokumenata javlja u nekoliko situacija. Najprije je to u okviru redovnoga poslovanja kad neki dokument ulazi u, primjerice, sustav za elektroničko

---

[3] Thurmair, G. *Making Term Extraction Tools Usable*. U: *Joint Conference combining 8th International Workshop of European Association for Machine Translation and 4th Controlled Language Applications Workshop*.Dublin :EAMT-CLAW, 2003.

[4] Smadja F., McKeown K. *Automatically extracting and representing collocations for language generation*. U: *Proceedings of the 28th Annual Meeting of the ACL*. 1990. Str. 252–259.

upravljanje dokumentima. Potrebno ga je digitalizirati i dodijeliti mu neku klasifikacijsku oznaku. Automatska identifikacija dokumenata važna je i u situaciji kad je neki dokument nastao u digitalnome obliku, bio otisnut, potpisan i kojega je tada potrebno digitalizirati i pohraniti u sustavu pismohrane. Treća situacija u kojoj je automatska identifikacija važna jest u trenutku kad se provodi digitalizacija postojećeg registraturnoga ili arhivskoga gradiva. Suvremeni skeneri koji automatski uvlače dokumente digitaliziraju velike količine gradiva. Takve se procese, koji se često provode u okviru projekata digitalizacije, naziva masovnom digitalizacijom. Upravo je u tim slučajevima potrebno omogućiti automatsku identifikaciju pojedinih dokumenata. Ponekad je u takvim procesima izazovno uopće automatski odrediti gdje neki dokument započinje, a gdje završava ili koji njegovi privitci zajedno s njime čine cjelinu, a kamoli još i odrediti gdje ga točno pospremiti u sustav digitalnoga repozitorija ili nekog drugog e-sustava. Upravo u tim trima situacijama jezična analiza sadržaja mogla bi učiniti pomak, jer je moguće pronalaziti jednoznačne semantički pune termine koji se pojavljuju u dokumentu, višerječne jedinice koje se uvijek pojavljuju jedna pored druge (sintagme – spojevi najmanje dviju punoznačnih riječi, kolokacije – višečlana jezična konstrukcija) i u međusobnim kombinacijama (detekcija uvijek istih, primjerice, četiriju kolokacija u istoj vrsti/klasi dokumenta) ili duže višerječne jedinice (npr. idiome i kolokacije) kojima se mogu koristiti za poboljšanje rada sustava.

U računalnoj ekstrakciji termina i kolokacija hrvatskoga jezika do sada su analizirani različiti pristupi u različite svrhe: statistički jezično neovisni pristupi (Seljan et al., u tisku), računalnojezični pristupi primjenom generičke metode,[5] lokalnih gramatika,[6] usporedna analiza hibridne metode i statističkih modela,[7]

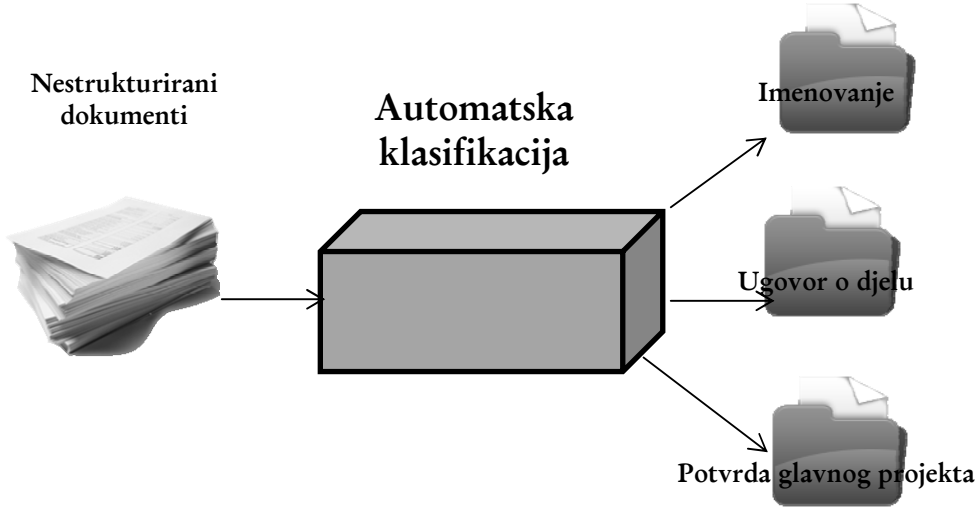
[5] Bekavac, B.,Tadić, M. *A Generic Method for Multi Word Extraction from Wikipedia*. U: *Proceedings of the 30th International Conference on Information Technology Interfaces*. Lužar-Stiffler, V., Hljuz Dobrić, V.; Bekić, Z. (ur.). Zagreb : SRCE University Computer Centre, University of Zagreb, 2008. Str. 663–667.

[6] Seljan, S., Gašpar, A.*First Steps in Term and Collocation Extraction from English-Croatian Corpus*. U: *Proceedings of 8th International Conference on Terminology and Artificial Intelligence*. Toulouse, 2009.

[7] Seljan, S., DalbeloBašić, B., Šnajder, J. Delač, D., Šamec-Gjurin, M., Crnec, D. *Comparative Analysis of Automatic Term and Collocation Extraction*. U: *The Future of Information Sciences: INFuture2009 – Digital Resources and Knowledge Sharing*. Stančić, H.,Seljan, S.,Bawden, D.,Lasić-Lazić, J.,Slavić, A. (ur.). Zagreb :Odsjekzainformacijskeznanosti, 2009. Str. 219–228.; Seljan, S.; Dunđer, I.; Gašpar, A.*From Digitisation Process to Terminological Digital Resources.* U: *Proceedings of the 36th International Convention MIPRO 2013*. Biljanović, P. (ur.). Rijeka : Croatian Society for Information and Communication Technology, Electronics and Microelectronics – MIPRO, 2013.

usporedba više različitih statističkih mjera,[8] ekstrakcija terminologije uz postupke sravnjivanja i statističke postupke za pronalaženje prijevoda.[9]

Iako postoje brojni radovi koji se odnose na ekstrakciju informacija, pojam kolokacija i termina nije jasno određen te često ovisi o svrsi sustava, vrsti korisnika i dostupnosti jezičnih resursa. Upravo zbog toga postoje brojni pristupi kojima se nastoje postići što bolji rezultati u ekstrakciji. Inicijalni rezultati ekstrakcije (tzv. „kandidati") uspoređuju se s prethodno izrađenom referentnom listom ili već postojećimstandardom (npr. EUROVOC, WordNet). Tu je riječ o ilustrativnom i jednostavnom primjeru, za koji čak i postoji obrazac, no mnogi dokumenti postoje kao privitci koji nisu u nekom standardnome obliku pa veliku poteškoću predstavlja identificiranje takvih dokumenata i njihova povezanost s osnovnim dokumentom. Za ilustraciju složenosti može poslužiti podatak dase u pojedinim tijelima državne i javne uprave razlikuje više od 300 različitih tipova (klasa i podklasa) dokumenata.[10]



Slika 1. Klasifikacija dokumenata korištenjem različitih pristupa identifikaciji nestrukturiranih dokumenata

[8] Petrović, S., Šnajder, J., Dalbelo Bašić, B., Kolar, M.Comparison of Collocation Extraction Measures for Document Indexing. *Journal of Computing and Information Technology.* 14(2006), str. 321–327.

[9] Ljubešić, N.,Vintar, Š., Fišer, D. *Multi-word term extraction from comparable corpora by combining contextual and constituent clues*. U: *Workshop on Building and Using Comparable Corpora (BUCC'12).* Istanbul, 2012.

[10] Primjer Agencije za lijekove i medicinske proizvode Republike Hrvatske (HALMED).

Jedan od pristupa automatskoj identifikaciji nestrukturiranih dokumenata je onaj korištenu AIDE projektu (Automatsko Indeksiranje DEskriptorima EUROVOC-a) nastalom kao rezultat suradnje između HIDRE[11], Fakulteta elektrotehnike i računarstva i Filozofskoga fakulteta Sveučilišta u Zagrebu. Cilj projekta AIDE bio je izraditi sustav za automatsko indeksiranje službenih tekstova na hrvatskome jeziku deskriptorima Pojmovnika EUROVOC. Automatsko označivanje primjenom kontroliranoga vokabulara rabi se za indeksiranje, pretraživanje i za uspostavljanje odnosa između sadržajno sličnih dokumenata. Pouliquen, Steinberger i Camelia[12] navode da samo ekstrakcija terminologije vrlo često nije dovoljna te da kod pravne dokumentacije dosiže odziv od svega 30,8%, zbog čega se uvode dodatne liste srodnih termina i dodatne statističke i jezične obrade teksta. No u tomje radu naglasak na postupcima ekstrakcije teksta.

U nastavku se objašnjavaju pristupi automatskoj ekstrakciji teksta koji bi se mogli iskoristiti u razvoju rješenja za (polu)automatsku identifikaciju i klasifikaciju poslovne dokumentacije odnosno identifikaciju registraturnoga i arhivskoga gradiva prilikom njihove digitalizacije. U ovome radu ti se pristupi sistematiziraju radi boljega pregleda nad cjelinom. Njihova podrobnija analiza i temeljitija međusobna usporedba zahtijevala bi mnogo više prostora te stoga prelazi opseg ovoga rada.

Pristupi automatskoj ekstrakciji

Danas postoje brojni alati koji provode ekstrakciju teksta. Svaki od alata najčešće ujedinjuje nekoliko različitih pristupa (npr. više različitih statističkih modela) koji se često kombiniraju s jezičnim modelima ili pojedinim *internetskim* jezičnim izvorima. Za rangiranje termina „kandidata" koriste se različitim statističkim mjerama, od kojih je najjednostavnija čestotnost pojavljivanja (engl. *frequency*) određenih različnica ili n-grama, zatim devijacija (engl. *deviation*), udaljenost (engl. *distance*) kojom se mjeri razlika između dvaju nizova ili skupova znakova, jačina ili značaj termina(engl. *strength*) itd. Primjerice, Levenshteinova udaljenost ukazuje na potreban broj umetanja, brisanja i zamjena znakova kako bi sedva niza znakova

[11] Danas Digitalni informacijsko-dokumentacijski ured Vlade Republike Hrvatske.

[12] Pouliquen, B, Steinberger R,Camelia I.*Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus.*U: *Proceedings of the Workshop Ontologies and Information Extraction at the Summer School : The Semantic Web and Language Technology – Its Potential and Practicalities (EUROLAN'2003).*Bukurešt, 2003.

izjednačila. Jaro-Winklerova distanca jedna je vrsta ponderiranog algoritma koja mjeri koliko je zajedničkih, tj. podudarajućih znakova, pri čemu se uzima u obzir mjesto pronalaska zajedničkih, tj. preklapajućih znakova. Može se koristiti matematičkom operacijom „presjek" i algoritmom za pretraživanje najvećeg zajedničkog podskupa znakova (engl. *longest common subsequence*). Nadalje, Jaccardov koeficijent mjeri sličnost odnosno različitost dvaju konačnih skupova znakova s pomoću matematičkih operacija „presjek" i „unija".Postoje brojne metrike primijenjene u algoritmima za ekstrakciju termina i kolokacija, kao primjericeC-vrijednosti,[13] NC-vrijednosti,[14] logaritamska vjerodostojnost/očekivanost (engl. *log-likelihood*) i uzajamna informacija/obavijesnost (engl. *mutual information*),[15] TF-IDF (engl. *term frequency-inverse document frequency*)[16] koja dodjeljuje veće vrijednosti ili veću „ključnost" terminima koji se češće javljaju u ulaznom dokumentu nego u cijelome korpusuitd. C-vrijednost proizlazi iz domenski neovisne metode za automatsko prepoznavanje termina kojom se efikasno razrješavaju ugniježđeni termini. Ona se temelji na frekventnosti termina „kandidata", frekventnosti termina „kandidata" koji su dio duljeg termina „kandidata", broju duljih termina „kandidata", duljini termina „kandidata" itd.[17] Potvrđivanje termina naknadno vrši domenski ekspert. NC-vrijednost proširenje je C-vrijednosti, a uzima se u obzir informacija koja proizlazi iz okoline termina, tj. tzv. kontekstni faktor.[18] Vjerodostojnost/očekivanost je funkcija koja obuhvaća parametre statističkog modela i opisuje funkciju parametara kada je poznat ishod nekog događaja. Logaritamska vjerodostojnost/očekivanost je logaritmirana funkcija vjerodostojnosti/očekivanosti, jer je na taj način jednostavnije raspoznavati vrijednosti funkcije. Uzajamna informacija/obavijesnost je mjera kojom se računa vjerojatnost supojavljivanja dviju nezavisnih varijabli, tj. različnica

[13] Ananiadou, S. *A methodology for automatic term recognition*. U:*Proceedings of 15th International Conference on Computational Linguistics COLING 94*. Kyoto, 1994. Str. 1034–1038.

[14] Frantzi, K, Ananiadou, S, Mima, H. Automatic Recognition of Multi-word Terms: the C-value/NC-value Method. *International Journal of Digital Libraries*. 3/2(2000), str. 115–130.

[15] Pantel, P, Lin, D. *A Statistical Corpus-Based Term Extractor*. U:*Advances in Artificial Intelligence : Advances in Artificial Intelligence 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2001, Ottawa, Canada, June 7-9, 2001: Proceedings*. Berlin: Springer, 2001. Str. 36–46.

[16] Basili, R, Moschitti, A, Pazienza, M, Zanzotto, F. *A contrastive approach to term extraction*.U: *Proceedings of 4th International Conference on Terminology and Artificial Intelligence (TIA 2001)*. Nancy, 2001. Str. 10.

[17] Frantzi, K., n. dj.

[18] Frantzi, K., n. dj.

uspoređena s vjerojatnošću njihovaodvojenog pojavljivanja. Ona dakle mjeri koliko poznavanje vjerojatnosti jedne varijable smanjuje neizvjesnost druge varijable.

Mnogi pristupi ekstrakciji terminologije kombiniraju različite metode, kao što su npr. logaritamska očekivanost za ekstrakciju terminologije iz jednojezičnog izvornog i ciljnog korpusa i frazno-temeljeno statističko strojno prevođenje (engl. *phrase-based machine translation*)[19] ili ekstrakcija iz jednojezičnog korpusa te naknadno sravnjivanje jedinica[20] ili kombiniranje nekoliko različitih statističkih jezično neovisnih metoda[21] ili generiranje termina „kandidata" iz sravnjenih jedinica te primjena frekvencije kako bi se odredila specifičnost i „težina" odnosno važnost termina.[22]

Jedan od prvih alata za ekstrakciju kolokacija[23]je Xtract, koji se koristi mjerom međusobne povezanosti (engl. *association measures* – AMs). Alat Collocate[24] je komercijalni alat koji se koristi točkastom procjenom uzajamne informacije (engl. *Pointwise Mutual Information*–PMI), koji omogućava ekstrakciju 12-grama i logaritamsku očekivanost (engl. *log-likelihood*) za ekstrakciju bigrama. Taj se alatne koristi morfološkom normalizacijom, kao što je lematizacija, ali može obrađivati POS (engl. *part-of-speech*) označeni korpus, u kojemu su jezične jedinice gramatički obilježene.

Seretan i Wehrli prikazuju alat za ekstrakciju kolokacija u računalno potpomognutom prevođenju temeljen na sintaktičkom pristupu koji se kombinira s mjerom međusobne povezanosti (AMs) koja pokazuje jačinu povezanosti između

---

[19] Haque, R, Penkale, S, Way, A. *Bilingual Termbank Creation via Log-Likelihood Comparison and Phrase-Based Statistical Machine Translation*. U: *Proceedings of 4th International Workshop on Computational Terminology (Computerm)*. 2014. Str. 42–51.

[20] Ha, L. A., Fernandez, G.,Mitkov, R.,Corpas, G. *Mutual Bilingual Terminology Extraction*. U: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Calzolari, N. et al. (ur.). European Language Resources Association (ELRA), 2008. Str. 1818–1824.

[21] Teixeira, L, Lopes, G, Ribeiro, R. A. *Automatic Extraction of Document Topics*. U: *Proceedings of DoCEIS'11 – Computing, Electrical and Industrial Systems 349*. IFIP Austria, 2011. Str. 101–108.; Teixeira, L, Lopes, G, Ribeiro, R. A.*Language Independent Extraction of Key Terms: An Extensive Comparison of Metrics. Agents and Artificial Intelligence*. Berlin &Heidelberg : Springer, 2013. Str. 69–82.

[22] Lefever, E, Macken, L, Hoste, V. *Language-independent bilingual terminology extraction from a multilingual parallel corpus*. U: *Proceedings of European Chapter of the Annual Meeting of the Association of Computational Linguistics Athens*.Atena, 2009. Str. 496–504.

[23] Smadja, F. (1991) i Smadja, F. (1993), n.dj.

[24] Barlow, M. Collocate 1.0: Locating collocations and terminology. TX:Athelstan, 2004.

riječi.[25] Alat TermeX razvijen na FER-u[26] koristi se širim spektrom mjere međusobne povezanosti (AMs) kako bi za 2-grame, 3-grame i 4-grame odabrao jednu od 14 mogućnosti AMs mjera (PMI, Dice koeficijent, logaritamska očekivanost, hi-kvadrat test) kombiniranu s računalnom jezičnom obradom.[27] Dice koeficijent je mjera pojmovno slična mjeri uzajamne informacije/obavijesnosti, ali s boljim rezultatima kod kolokacija niske frekvencije. TermeX alat koristi se morfološkom normalizacijom, POS označavanjem i filtriranjem prema frekvencijama.

Goldman temelji ekstrakciju terminologije na sintaktičkom parseru, tj. sintaksnom analizatoru.[28] Wherli, Seretan i Nerima prikazuju primjer hibridnog sustava.[29] ITS-2 sustav temelji se na detaljnoj jezičnoj analizi parsera i korištenju jednojezičnog leksikona. Primjenom odgovarajućeg modela u sustavu transfera stvara se predikatno-argumentna struktura s detaljnim prikazom informacija te se identificiraju višerječne jedinice i kolokacije. Ekstrakcija se provodi primjenom hibridnog sustava u kojem se kombiniraju jezične informacije i statističke metode.Wu analizira ekstrakciju informacija iz dvojezičnog korpusa primjenom logaritamske očekivanosti i postupaka sravnjivanja.[30]

Postoje brojni alati koji funkcioniraju kao samostalne internetske aplikacije, koje se najčešće temelje na jednom ili više statističkih pristupa uz mogućnost dodatnog filtriranja termina „kandidata", određivanju minimalne frekvencije termina „kandidata" i duljine češćih sintagmi, mogućnosti povezivanja s nekom ontologijom ili,primjerice, naizračunu C-vrijednosti koja obuhvaća statističku i lingvističku analizu,kao što suPOS označavanje korpusa, lingvističko filtriranje tipova

---

[25] Seretan, V., Nerima, L.,Wehrli, E.*A tool for multi-word collocation extraction and visualization in multilingual corpora*. U: *Proceedings of the Eleventh EURALEX International Congress*. Lorient, 2004. Str. 755–766.

[26] Delač, D., Krleža, Z., Dalbelo B., Bojana, Šnajder, J., Šarić, F. *TermeX: A Tool for Collocation Extraction. Lecture Notes in Computer Science(Computational Linguistics and Intelligent Text Processing)*. 2009. Str. 149–157.

[27] Detaljno prikazano u Petrović, S. et al., n. dj.

[28] Goldman, J-P., Wehrli, E. *FipsCo: A syntax-based system for terminology extraction*. U: *Grammar and Natural Language Processing Conference*. Université du Quebec at Montreal, 2001.

[29] Wehrli, E., Seretan, V., Nerima, L., Russo, L. *Collocations in a rule-based MT system: A case study evaluation of their translation adequacy*. U: *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation*. Barcelona 2009. Str. 128–135.

[30] Wu C-C, Chang J. S.Bilingual collocation extraction based on syntactic and statistical analyses. *Computational Linguistics and Chinese Language Processing*. 9(2004), str. 1–20.

ekstrahiranih termina te uporaba popisa tzv. stop-riječi[31]s većim naglaskom na statističkoj analizi.

Zaključak

Postupci ekstrakcije terminologije i kolokacija tek su jedan korak u mogućem modelu poluautomatske ili automatske identifikacije i klasifikacije gradiva. Iako u ekstrakciji postoje različiti pristupi, ne postoji preporučeni model koji bi davao najbolje rezultate. Hibridni modeli najčešće kombiniraju različite metode čiji odabir ponekad ovisi o krajnjem korisniku i upotrebi ekstrahiranih termina. Postupci automatske klasifikacije zahtijevaju treniranje sustava odnosno„učenje“ na velikoj količini provjerene dokumentacije, tj. dokumentacije za koju je unaprijed poznato kako je klasificirana i kojim je ključnim riječima obilježena. S obzirom na nestandardizirani oblik dokumenata i na činjenicu da pojedini pojmovi nisu jasno izraženi niti je unaprijed poznato mjesto njihova pojavljivanja u tekstu, potrebni su dodatni procesi za sadržajno povezivanje, analizu i računalnu obradu. S obzirom na složenost postupka, raznolikost poslovne dokumentacije i arhivskoga gradiva u strukturi, sadržaju, a ponekad i jeziku, trenutačno je potrebna značajna ljudska intervencija u svim fazama rada.

Postupci automatske ekstrakcije teksta, kojima se primarno koristi u području računalne obrade prirodnoga jezika, pokazuju potencijal za korištenje i u području arhivske struke u kontekstu identifikacije i klasifikacije gradiva. Njima bi se podjednako moglo koristiti za analizu digitaliziranoga gradiva, koje je uspješno obrađeno OCR programom, kao i gradiva koje je izvorno nastalo u digitalnome obliku. Ipak, za konkretno i kvalitetno rješenje namijenjeno nestrukturiranim dokumentima potrebno je provesti dodatna istraživanja i testiranja. Konačno, važno je naglasiti da je tehnologija dovoljno uznapredovala da bi se uskoro i u području arhivske struke mogla početi pojavljivati rješenja koja se u pozadini koriste principima i pristupima sistematiziranima u ovome radu. Zbog toga ih je potrebno razumjeti, jer kad se rješenja jednom počnu pojavljivati, bit će važno procijeniti njihovu kvalitetu i iskoristivost.

[31] Frantzi, K. et al., n. dj.

Iz provedene analize trenutačnoga stanja te uočavanja problema identifikacije i automatske klasifikacije nestrukturiranih dokumenata, kojemu trenutno razvijena programska rješenja teško ili gotovo nikako ne mogu odgovoriti, teprepoznavanja područja strojnoga prevođenja kao područja iz kojeg bi mogla doći nova rješenjakao i sistematizacije principa i pristupa koji se u njemu rabe, jasno je da se u arhivskoj struci nazire jedan novi smjer kretanja te otvara prostor za nova istraživanja.

Literatura

1. Ananiadou, S. *A methodology for automatic term recognition*. U:*Proceedings of 15th International Conference on Computational Linguistics COLING 94*. Kyoto, 1994. Str. 1034–1038.
2. Barlow, M.*Collocate 1.0: Locating collocations and terminology*. TX:Athelstan, 2004.
3. Basili, R., Moschitti, A., Pazienza, M., Zanzotto, F. *A contrastive approach to term extraction*. U:*Proceedings of 4th International Conference on Terminology and Artificial Intelligence (TIA 2001)*. Nancy, 2001. Str. 10.
4. Bekavac, B., Tadić, M. *A Generic Method for Multi Word Extraction from Wikipedia*. U: *Proceedings of the 30th International Conference on Information Technology Interfaces*. Lužar-Stiffler, V., Hljuz Dobrić, V., Bekić, Z. (ur.). Zagreb : SRCE University Computer Centre, University of Zagreb, 2008. Str. 663–667.
5. Daille, B., Gaussier, E., Lang´e, J-M. *Towards automatic extraction of monolingual and bilingual terminology*. U: *Proceedings of 15 Conference on Computational Linguistics COLING*. Kyoto, 1994. Str. 515–521.
6. Delač, D., Krleža, Z., DalbeloBašić, B., Šnajder, J., Šarić, F. *TermeX: A Tool for Collocation Extraction. Lecture Notes in Computer Science (Computational Linguistics and Intelligent Text Processing)*. 2009. Str. 149–157.
7. Frantzi, K., Ananiadou, S., Mima, H. Automatic Recognition of Multi-word Terms: the C-value/NC-value Method. *International Journal of Digital Libraries*. 3/2(2000), str. 115–130.

8. Goldman, J-P., Wehrli, E.*FipsCo: A syntax-based system for terminology extraction.Grammar and Natural Language Processing Conference*. Université du Quebec at Montreal, 2001.
9. Ha, L. A., Fernandez, G., Mitkov, R., Corpas, G. *Mutual Bilingual Terminology Extraction*. U: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.Calzolari, N. et al. (ur.).European Language Resources Association (ELRA), 2008. Str. 1818–1824.
10. Haque, R., Penkale, S., Way, A. *Bilingual Termbank Creation via Log-Likelihood Comparison and Phrase-Based Statistical Machine Translation*. U:*Proceedings of 4th International Workshop on Computational Terminology (Computerm)*, 2014. Str. 42–51.
11. Lefever, E., Macken, L., Hoste, V. *Language-independent bilingual terminology extraction from a multilingual parallel corpus*. U:*Proceedings of European Chapter of the Annual Meeting of the Association of Computational Linguistics Athens*.Atena, 2009. Str. 496–504.
12. Ljubešić, N., Vintar, Š., Fišer, D. *Multi-word term extraction from comparable corpora by combining contextual and constituent clues*. U: *Workshop on Building and Using Comparable Corpora (BUCC'12)*. Istanbul, 2012.
13. Pantel, P., Lin, D. *A Statistical Corpus-Based Term Extractor*. U:*Advances in Artificial Intelligence : Advances in Artificial Intelligence 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2001, Ottawa, Canada, June 7-9, 2001: Proceedings*. Berlin:Springer, 2001. Str. 36–46.
14. Petrović, S., Šnajder, J., Dalbelo Bašić, B., Kolar, M.Comparison of Collocation Extraction Measures for Document Indexing. *Journal of Computing and Information Technology*. 14(2006), str. 321–327.
15. Pouliquen, B., Steinberger R., Camelia I. *Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus*. U:*Proceedings of the Workshop Ontologies and Information Extraction at the Summer School: The Semantic Web and Language Technology – Its Potential and Practicalities (EUROLAN'2003)*. Bukurešt, 2003.

16. Quigley, R. Big data in commodity markets. Business Opportunity or Another Fad? *Commodities Now*, ožujak 2014, str. 61–65. URL: http://www.datagenicgroup.com/wp-content/uploads/2014/06/Big_Data_In_Commodity_Markets.pdf (3.5.2015.).
17. Seljan, S., Dalbelo Bašić, B., Šnajder, J., Delač, D., Šamec-Gjurin, M., Crnec, D. *Comparative Analysis of Automatic Term and Collocation Extraction*. U: *The Future of Information Sciences: INFuture2009 – Digital Resources and Knowledge Sharing*. Stančić, H., Seljan, S., Bawden, D., Lasić-Lazić, J., Slavić, A. (ur.). Zagreb :Odsjek za informacijske znanosti, 2009. Str. 219–228.
18. Seljan, S., Dunđer, I., Gašpar, A. *From Digitisation Process to Terminological Digital Resources*.U: *Proceedings of the 36th International Convention MIPRO 2013*. Biljanović, P. (ur.). Rijeka : Croatian Society for Information and Communication Technology, Electronics and Microelectronics – MIPRO, 2013.
19. Seljan, S.,Gašpar, A. *First Steps in Term and Collocation Extraction from English-Croatian Corpus*.U: *Proceedings of 8th International Conference on Terminology and Artificial Intelligence*. Toulouse, Francuska, 2009.
20. Seljan, S., Dunđer, I., Stančić, H. *Extracting Terminology by Language Independent Methods*. [U tisku.]
21. Seretan, V., Nerima, L., Wehrli, E. *A tool for multi-word collocation extraction and visualization in multilingual corpora*. U: *Proceedings of the Eleventh EURALEX International Congress*. Lorient, 2004. Str. 755–766.
22. Smadja F., McKeown K. *Automatically extracting and representing collocations for language generation*. U: *Proceedings of the 28th Annual Meeting of the ACL*, 1990. Str. 252–259.
23. Smadja, F. *From n-grams to collocations : An evaluation of Xtract*. U: *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics*, 1991. Str. 279–284.
24. Smadja, F. *Retrieving collocations from text: Xtract*. U:*Proceedings of 31th Annual Meeting of the Association for Computational Linguistics*. 19(1993), str. 143–177.

25. *Technologies for the Processing and Retrieval of Semi-Structured Documents – Experience from the CADIAL project*.Tadić, M., DalbeloBašić, B., Moens, M-F. (ur.).Zagreb :Croatian Language Technologies Society, 2009.
26. Teixeira, L., Lopes, G., Ribeiro, R.A. *Automatic Extraction of Document Topics. Proceedings of DoCEIS'11 – Computing, Electrical and Industrial Systems 349*. IFIP Austria, 2011. Str. 101–108.
27. Teixeira, L., Lopes, G., Ribeiro, R.A. *Language Independent Extraction of Key Terms: An Extensive Comparison of Metrics. Agents and Artificial Intelligence*. Berlin &Heidelberg :Springer, 2013. Str. 69–82.
28. Thurmair, G. *Making Term Extraction Tools Usable*. U:*Joint Conference combining 8th International Workshop of European Association for Machine Translation and 4th Controlled Language Applications Workshop*. EAMT-CLAW, 2003.
29. Wehrli, E., Seretan, V., Nerima, L., Russo, L. *Collocations in a rule-based MT system: A case study evaluation of their translation adequacy*. U:*Proceedings of the 13th Annual Meeting of the European Association for Machine Translation*. Barcelona, 2009. Str. 128–135.
30. Wu C-C., Chang J. S.Bilingual collocation extraction based on syntactic and statistical analyses. *Computational Linguistics and Chinese Language Processing*. 9(1) (2004), str. 1–20.

Summary

THE CONCEPT OF THE AUTOMATIC CLASSIFICATION OF THE REGISTRY AND ARCHIVAL RECORDS

Electronic Document and Records Management Systems (EDRMS) that are most often parts of the more total Electronic Corporate Management Systems (ECMS) usually encompass documents and records that are originally created in the digital form, as well as the non-digitalised documents. While the originally digitized records are fairly easy to describe during their creation and add to them all the necessary metadata, the problems occur with those records that enter the system while undergoing the digitization process. If

it concerns the large amount of data, in which process the documents are diverse and do not have some unique or recurring characteristics, it is not easy to determine the document in question or to classify it correctly and add metadata to it. The authors analyse and show the possibilities of solutions that belong to the area of the statistically based lingual technologies and they also research their possible application in the area of (semi)automatic classification of the registry and archival records. This paper explains the basic starting points of individual methods, the possibilities of the automatic extraction of the text, the methods of statistical processing and laying the foundation for the (semi)automatic classification. The authors demonstrate results of testing the applied methods on the specific archival records and draw conclusions about the possible future research directions.

Keywords: *automatic classification, computer processing of natural language, statistical methods, digitization, archival records*

Translated by Marijan Bosnar