Зборник радова Факултета техничких наука, Нови Сад

UDK: 004.9
DOI: https://doi.org/10.24867/20BE12Rokvic

РАЗВОЈ ОНТОЛОГИЈЕ УПОТРЕБОМ МЕТОДА ОБРАДЕ ПРИРОДНОГ ЈЕЗИКА

ONTOLOGY BUILDING USING NATURAL LANGUAGE PROCESSING METHODS

Невена Роквић, *Факултет техничких наука , Нови Сад*

Област – ЕЛЕКТРОТЕХНИКА И РАЧУНАРСТВО

Кратак садржај – *Онтологије се користе широм различитих области као супериорно решење за приказивање доменског знања. Традиционални системи грађења онтологије подразумевају комплексан процес који захтева велике количине ресурса и времена. Са развојем науке и области обраде природног језика тај процес је могуће делимично или у потпуности аутоматизовати. Овај рад тежи да применом машинског учења попуни онтологију из домена рачунарских наука. Као улазни подаци за тренирање неуронске мреже коришћени су наставни материјали, где су као резултат добијене уградње речи којима је аутоматски попуњавана онтологија. Добијени модел је евалуиран поређењем са оригиналним подацима из наставних материјала.*

Кључне речи: *Онтологија, Обрада природног језика, Неуронске мреже, Уградња речи*

Abstract – *Ontologies are widely used in numerous areas as a superior solution for domain knowledge representation. Traditional ontology building systems include many complicated and time-consuming tasks, but with the development of science it is possible to automate the process, partially or entirely. This paper has a goal to represent the process of computer science ontology building using machine learning. Neural network was trained on university courses data and as a result word embeddings were used to build the ontology. The model was evaluated by comparing word embeddings with the ground truth.*

Keywords: *Ontology, Natural Language Processing, Neural Networks, Word Embeddings*

1. УВОД

У домену школовања онтологија се употребљава као алат за интеграцију великих скупова података истраживачких радова, информација добијених из истраживачких радова, итд. Овакве онтологије служе као помоћ за разумевање на који начин функционише динамика истраживачког процеса, за класификацију публикација, предвиђање трендова, итд. За разлику од других истраживачких области које имају богато описане атоматски генерисане онтологије, онтологије из области рачунарских наука су се развијале знатно спорије [1].

У систему који је описан у овом раду представљен је процес аутоматског грађења онтологије из домена рачунарских наука помоћу курсева са Института за технологију из Масачусетса (*MIT*-а), где су коришћени слајдови са предавања из 10 различитих области покривених наставним материјалима.

Главна идеја за имплементацију овог система лежи у томе да се употребом машинског учења аутоматизује процес попуњавања онтологије, као и да се провери квалитет добијених речи и њихових веза. Онтологија је формирана тренирањем плитке неуронске мреже *word2vec* [2] за коју је као улазни скуп података служио корпус који садржи стотине слајдова са теоријским и практичним садржајем курсева из рачунарских наука са *MIT*-а. Након прикупљања података извршено је детаљно процесуирање података, што подразумева примену техника из области обраде природног језика као што су чишћење, лематизација, токенизација, итд. Добијене речи и њихове уграђене речи (енг. *word embeddings*) су затим употребљени за попуњавање онтологије. Онтологија садржи две главне класе, и свака од њих три подкласе. Класа која сачи састав курсева је детаљно попуњена садржајем курсева, тачније најчешће појављиваним речима и њиховим сличним појмовима као подкласама. Остале класе представљају врсту мета-података о онтологији, и оне нису попуњаване због временске захтевности тог процеса. Мали део онтологије је ручно попуњаван, као што су области којима су попуњаване речи и везе између речи. Атрибути речи су додати аутоматски, јер се могу директно добити из тренираног *word2vec* модела. Они се пре свега односе на проценат сличности речи између себе и на позицију речи у листи најфреквентнијих речи из корпуса.

2. ПРЕТХОДНА РЕШЕЊА

Појам онтологије је саставни део области Семантичког веба дужи низ година. Будући да је ручно грађење онтологије временски захтевно, у све више радова се предлаже аутоматско попуњавање онтологије. Пратећи трендове науке научници са Универзитета у Бечу предлажу систем за учење онтологије који као улазне податке узима доменски текст, *Wordnet*, *DBPedia*-u и подаци са одређених друшвених медија. Из прикупљеног корпуса систем извлачи битне појмове и везе између појмова. Грађење онтологије започиње од самог почетка тако што користи неколико улазних појмова, и на основу њих генерише нови појам и потенцијалне повезане кандидате. Наредни корак је интеграција свих кандидата у семантичку мрежу и проналажење 25

НАПОМЕНА:
Овај рад проистекао је из мастер рада чији ментор је био др Милан Сегединац, ванр. проф.

најбитнијих кандидата за концепт помоћу неуронске мреже. На крају се врши евалуација добијених појмова и интеграција добијених појмова у онтологију, чиме се добија проширена верзија полазне онтологије. Та проширена онтологија је полазна тачка за даље генерисање појмова и повратак на први корак овог циклуса. Овај систем су поредили са две варијације *word2vec* модела (униграм - модел трениран на појединачним речима и биграм - модел трениран на појединачним речима и биграмима). Најбољи резултат је показао *word2vec* модел базиран на појединачним речима. Закључили су да са повећањем броја концепата опада квалитет новогенерисаних кандидата, док је са *word2vec* моделом обрнута ситуација [3].

Једна од примена онтологија у образовању је у склопу система препоруке. Будући да је одабир факултета често захтеван и стресан процес, средњошколцима је преко потребно вођење и подршка. Због затрпаности информацијама о различитим студијским програмима неопходно је ефикасно процесуриати информације да би се дошло до решења. У овом раду је представљен систем за препоруку заснован на онтологији који је развијен помоћу техника машинског учења. Направљен је да препозна јаче и слабије стране сваког студента, као и његова интересовања и могућности. Главне компоненте овог система баве се препоруком универзитета ђацима на основу образовног пута студената који поседују минимално диплому основних студија. На основу података о њиховим интересовањима током средње школе, изабраним студијским програмом и доменом у којем су тренутно запослени систем учи о њиховом образовном путу. Главни закључак овог рада је да је грађење ефикасног система за препоруку који се заснива на технологијама семантичког веба захтева дугорочну посвећеност и постојање једноставних алата који би омогућили креирање, објављивање и добављање семантичких података [4].

3. МЕТОДОЛОГИЈА

У наредном тексту детаљно су појашњени скуп података, архитектура система, процес тренирања модела и попуњавање онтологије добијеним подацима.

3.1. Скуп података

Скуп података који се користи у овом раду је формиран од материјала преузетог са интернет странице [5] која садржи сав наставни материјал са *MIT*-а. Иако на самој страници постоји мноштво различитих материјала различитог формата као што су задаци, слике, слајдови и сл., за потребе овог пројекта одабрани су само наставни слајдови, будући да они садрже највећу количину текста од свих претходно наведених извора. На слици 1 илустрован је процес прикупљања података помоћу којег се добијају подаци спремни за чување у неком од формата и даљу обраду. У овом случају подаци су сачувани у текстуалне фајлове, подељени по областима. Након тога, подаци су претпроцесуирани коришћењем разних техника за обраду природног језика (као што су уклањање специјалних карактера, стоп речи и лематизација) и претворени у токене, који су даље служили као улазни корпус за тренирање неуронске мреже.

3.2. Архитектура система

Будући да је корпус подељен према наставним областима, за сваку од области је трениран *word2vec* [2] модел. Он се базира на плиткој неуронској мрежи, и има две варијације: *CBOW* (Continuos Bag-Of-Words) (slika 2) и *skip-gram*. Главна разлика између ова два модела је у томе сто се код *CBOW* модела као инпут користи контекст да би се предвидела циљана реч, а код *skip-gram* модела је обрнуто: користи се нека реч да би се предвидео њен контекст. У овом раду коришћена је *CBOW* варијанта.

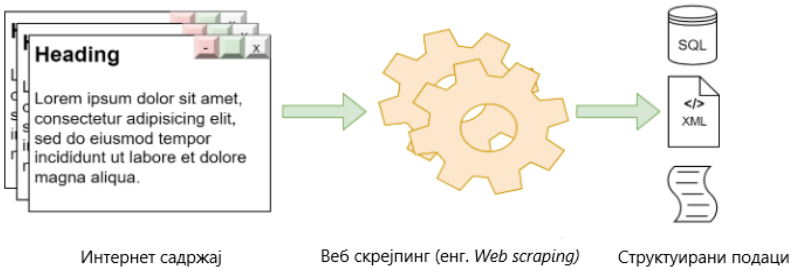
Тренирањем модела добијене су уградње речи (енг. *word embeddigs*), којима је попуњавана онтологија коришћењем пакета програмског језика *Python* [6] и *Protégé* [7] алата.

3.3. Попуњавање онтологије

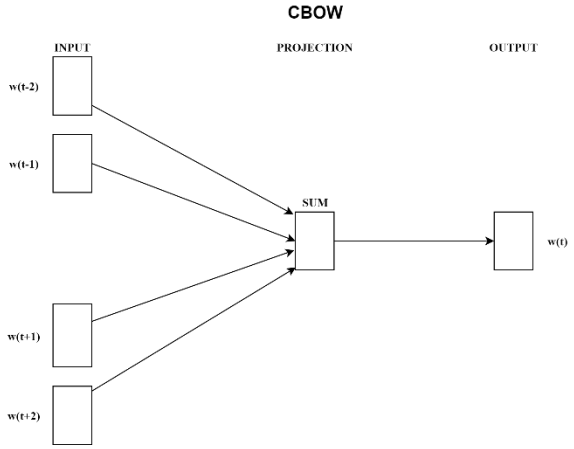
Изградња онотологије од уградњи речи добијених тренирањем *word2vec* модела састоји се из неколико фаза. Прва фаза подразумева ручно одређивање главних појмова (класа) које ће служити као база за остатак. Будући да је циљ ове онтологије приказ појмова из области рачунарства као главне класе издвојили су се следећи појмови:

- Topic Content
- ComputerScienceModuleContent
- ExamExample
- StudyMaterial
- LectureNotes
- SimilarTerm,
- FrequentTerm
- Slides
- StudyLevel

Ове класе су додате ручно уз помоћ *Protégé* [7] алата, і и онтологија је сачувана у *.owl* формату.



Слика 1. *Прикупљање података*

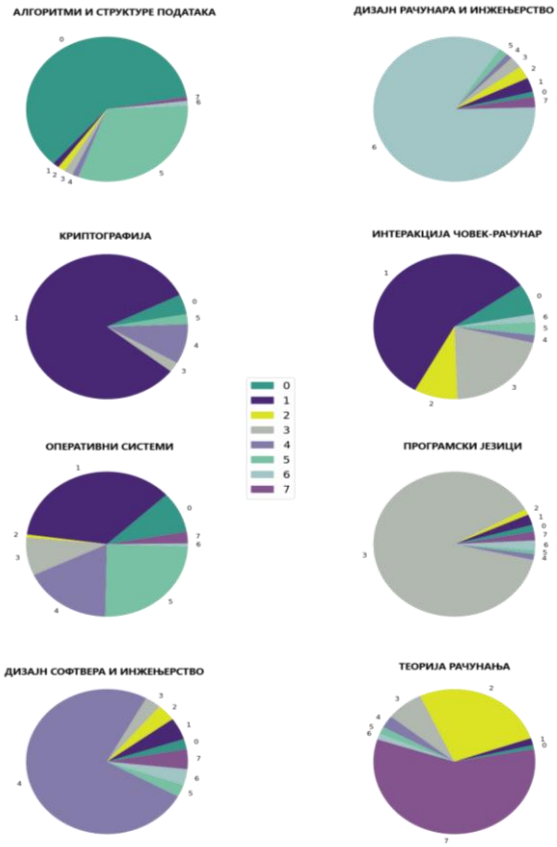


Слика 2. *word2vec (CBOW)*

4. РЕЗУЛТАТИ

Задатак овог рада спецефичан је по томе што не постоји велика литература која се бави попуњавањем онтологије из области рачунарских наука на овај начин. Имајући то на уму, као начин евалуације квалитета решења узета је кластеризација уградњи речи и поређење припадности уградњи кластерима са њиховом оригиналном класом (области).

На слици 3 приказане су појединачне области и удео сваког од кластера унутар њих.



Слика 3. Кластери-евалуација

На основу резултата добијених кластеризацијом види се да већина области највећим делом припада једном кластеру. На пример, област Алгоритми и структуре података највећим делом припада кластерима 0 и 5, а Оперативни системи такође великим делом припадају кластеру 5, што се може приписати сличности појмова који се налазе у обе области. Област Дизајн рачунара и инжињерство има врло јасну припадност кластеру 6 са убедљиво највећим уделом, и врло малим уделима осталих 6 области. Област која садржи највише ‘расутих’ појмова, то јест спада са сличним процентом у више кластера је област Оперативни системи.

Тумачењем анализе кластера закључује се да постоје сличности између кластера добијених од стране уградњи речи и оригиналних лабела које су ручно одређене на почетку, што говори да ово решење може бити корисно за изградњу онтологије из области рачунарства.

5. ЗАКЉУЧАК И БУДУЋА ИСТРАЖИВАЊА

У овом раду је представљен процес конструисања онтологије коришћењем машинског учења, од прикупљања података и њихове припреме, преко тренинрања модела заснованих на плиткој неуронској мрежи до попуњавања онтологије коришћењем добијеног садржаја. Приказани су изазови приликом процеса прикупљања неструктуираних података, разлике између садржаја различитих области и конструисање онтологије од самог почетка комбиновањем традиционалног ручног попуњавања са аутоматским, користећи *Python* пакет за конструисање онтологије.

Показано је у пракси да се коришћењем јавно доступног материјала може изградити онтологија широке примене у области образовања, која може бити од велике користи и студентима и професорима. Студентима може користити као помоћ при одабиру правца кретања студија, а такође може служити као асистенција током праћења наставе тиме што ће уз помоћ семантички богате онтологије моћи да посматрају међусобну повезаност области и информације везане за предмете које прати. Са друге стране, једна од користи за професоре јесте употреба за прилагођавање наставног програма и садржаја наставе у зависности од повезаности дате области са другим областима, а и на основу других потенцијалних везаних мета-података о тој области (нпр. броја пријављених студената, њихових интересовања, итд.). Такође, још једна област у којој се може применити овај тип онтологије јесте код индивидуалног праћења развоја студената и његових склоности.

Један од главних изазова приликом израде овог рада јесте чињеница да се у овом раду користе искључиво слајдови са предавања. Они по својој природи у највећем броју случајева садрже скраћене лекције и непотпун садржај, па и квалитет решења варира од области до области. Области које садрже детаљније описе су понудиле боље решење, док области које већински садрже скраћенице и формуле немају богат речник и самим тим опада квалитет тог дела онтологије. Управо због тога један од праваца будућих истраживања би било унапређивање онтологије проширењем корпуса (додавањем аудио и видео лекција, примера испита, белешки са вежби, белешки са предавања, итд) ради побољшања квалитета постојеће онтологије.

6. ЛИТЕРАТУРА

[1] Salatino, Angelo A., et al. "The computer science ontology: A comprehensive automatically-generated taxonomy of research areas." Data Intelligence 2.3 (2020): 379-416.

[2] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[3] Wohlgenannt, Gerhard, and Filip Minic. "Using word2vec to Build a Simple Ontology Learning System." International Semantic Web Conference (Posters & Demos). 2016.

[4] Obeid, Charbel, et al. "Ontology-based recommender system in higher education." Companion Proceedings of the The Web Conference 2018. 2018.

[5] https://ocw.mit.edu/index.htm.(приступљено у јуну 2022.)

[6] vanRossum, Guido. "Python reference manual." Department of Computer Science [CS] R 9525 (1995).

[7] Noy, N. F., Crubézy, M., Fergerson, R. W., Knublauch, H., Tu, S. W., Vendetti, J., & Musen, M. A. (2003). Protégé-2000: an open-source ontology-development and knowledge-acquisition environment<. AMIA Symposium, (str. 953-953).

Кратка биографија:

Невена Роквић рођена је у Новом Саду 1994. године. Дипломски рад на на Факултету техничких наука из области Електротехнике и рачунарства – Примењене рачунарске науке одбранила је 2019. године. Мастер студије на Факултету техничких наука из области Електротехнике и рачунарства – Интелигентни системи уписује исте године.
контакт: nevrokvic@gmail.com