Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.896
DOI: https://doi.org/10.24867/28BE29Ivanovic

СУМАРИЗАЦИЈА НАУЧНИХ РАДОВА НА СРПСКОМ ЈЕЗИКУ ПРИМЕНОМ NLP МЕТОДА

SUMMARIZATION OF SCIENTIFIC PAPERS IN SERBIAN LANGUAGE USING NLP METHODS

Наташа Ивановић, *Факултет техничких наука, Нови Сад*

Област – ЕЛЕКТРОТЕХНИКА И РАЧУНАРСТВО

Кратак садржај – *У раду је представљен систем за сумаризацију научних радова на српском језику применом NLP метода са циљем да се посао истраживача олакша кроз аутоматско генерисање апстракта. Решење је имплементирано кроз два модула – фаза екстракције (TextRank алгоритам) и фаза апстракције (Sequence-to-Sequence модели), где се трансформер модел показао као најбољи избор.*

Кључне речи: *Аутоматска сумаризација текста, NLP, Sequence-to-Sequence модели, Трансформер модели, TextRank алгоритам*

Abstract – *The paper presents a system for summarizing scientific papers in Serbian using NLP methods, aiming to facilitate researchers' work through automatic abstract generation. The solution is implemented in two modules – the extraction phase (TextRank algorithm) and the abstraction phase (Sequence-to-sequence models), for which the Transformer model has proven to be the best choice.*

Keywords: *Automatic text summarization, NLP, Sequence-to-Sequence models, Transformer models, TextRank algorithm*

1. УВОД

Системи за сумаризацију текста представљају интегрални део различитих решења у оквиру обраде природног језика (*NLP* – енгл. *Natural Language Processing*). Задатак сумаризације текста има за циљ очување кључних информација и контекста, редукујући оригинални текст на краћу верзију [1]. Овај рад се фокусира на генерисање сажетака научних радова. Апстракт представља кратак опис научног рада који садржи кључне информације о истраживању [2] - има значајну улогу, јер истраживачи често на основу садржаја апстракта одлучују да ли ће наставити са читањем остатка рада.

Систем за генерисање апстракта има потенцијал да олакша посао истраживача и помогне у ефикаснијем и бржем комуницирању кључних аспеката рада. Аутоматски генерисани апстракти могу пружити објективнији приказ рада и побољшати конзистентност у стилу и садржају [2].

У овом раду представљено је решење за аутоматско генерисање апстракта на српском језику применом екстракције и апстракције текста. У првом кораку користи се *TextRank* алгоритам [3] за копирање, односно екстракцију најбитнијих реченица из рада, које служе као улаз у фазу апстракције. У сврхе апстракције користе се неуронске мреже засноване на енкодер-декодер архитектури [4], у циљу генерисања нових фраза које нису обавезно део оригиналног текста. У наставку биће анализирано стање у области, а затим ће бити дат преглед методологије, извршених експеримената и њихових резултата.

2. СТАЊЕ У ОБЛАСТИ

Интересовање за аутоматску сумаризацију текста датира још из половине 20. века. Ручно писање апстракта подложно је субјективној процени истраживача – аутори рада [5] предлажу решење које се заснива на копирању најзначајнијих реченица рангираних на основу броја понављања често кориштених речи и кључних речи, речи у оквиру наслова и поднаслова, као и на основу саме позиција реченице у тексту. Старији приступи ослањају се на хеуристику и прецизно дефинисане параметре. У научним радовима се исте информације формулишу на различите начине у зависности од поглавља. Тешко је избећи дупликате у генерисаним апстрактима примењујући овакав приступ.

У раду [6] представљена је примена *TextRank* алгоритма у сврхе екстрактивне сумаризације текста писаног на српском језику. Аутори истичу да су досадашња истраживања у овом домену већински рађена над текстовима на енглеском и да је стога овај проблем изазован за решавање. Алгоритам рачуна косинусну сличност међу реченицама и генерише матрицу сличности, која се затим конвертује у граф, где су реченице чворови, а гране скорови сличности. Финални сажетак представља топ три рангиране реченице. Решење је евалуирано применом *ROUGE* метрике.

Рани приступи апстрактној сумаризацији укључивали су (1) *sentence compression*, чији циљ је да редукује дужину реченице уклањајући садржај који није есенцијалан; (2) *sentence fusion* где се комбинује неколико независних реченица у једну кохерентну целину и (3) *sentence revision* током ког се генеришу реченице које нису део оригиналног текста [7]. Данас се за потребе апстракције користе *sequence-to-*

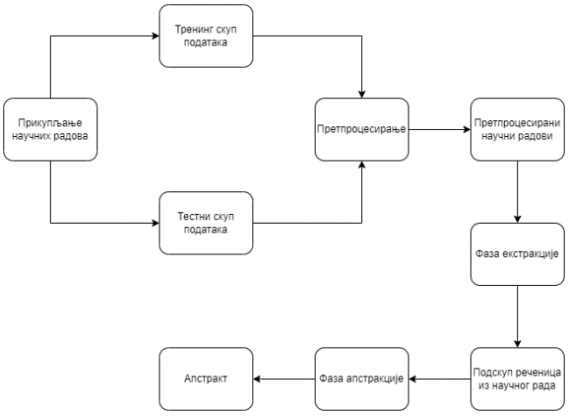
НАПОМЕНА:
Овај рад проистекао је из мастер рада чији ментор је био др Александар Ковачевић, ред. проф.

sequence модели са енкодер-декодер архитектуром. Енкодери и декодери су најчешће имплементирани као *RNN* (енгл. *Recurrent Neural Network*). Енкодовање дугачких секвенци и даље је отворени проблем за ове моделе. Архитектура је унапређена увођењем *attention* механизма. Кључна идеја је идентификација значајнијих речи и фраза у документу и прослеђивање те информације декодеру [7].

У раду [8] представљен је комбиновани приступ сумаризацији текста, уз нагласак да овакав приступ значајно унапређује добијене резултате. Како би се избегао проблем предугачких докумената, примењује се корак екстракције. Користе се два модела који се заснивају на енкодер-декодер архитектури. Код првог модела се за енкодер користе два *Bi-LSTM*-а, а за декодер један *LSTM*. За други модел, користе се два *LSTM*-а. За потребе апстракције кориштен је *GPT* трансформер модел. Кориштена су два јавно доступна скупа података: *arXiv* и *PubMed*. Кориштена је *ROUGE* метрика за евалуацију. Комбиновани приступ аутоматској сумаризацији текста уз коришћење трансформер модела доноси обећавајуће резултате.

3. МЕТОДОЛОГИЈА

Сумаризација текста одвија се у две фазе. Прво се примењују екстрактивне методе за издвајање најзначајнијих реченица, а затим се тај међу-резултат прослеђује у фазу апстракције. Улаз у систем представља научни рад, а излаз из система представља апстракт. Слика 3.1 приказује груб преглед архитектуре решења. У наставку овог поглавља биће детаљније објашњени кораци обраде научних радова и модули система.



Слика 3.1. *Преглед архитектуре решења*

3.1. Прикупљање података

Скуп научних радова писаних на српском језику преузет је скрејповањем званичне *web* странице Зборника радова Факултета техничких наука[1] помоћу *selenium* библиотеке. Разлог за овај одабир лежи у томе што већина радова прати предефинисан формат. Иницијално је прикупљено 2335 радова. Након анализе њихове структуре и форматирања, избачени су они који су одскакали од стандарда. После чишћења скупа података преостало је укупно 2215 радова који су подвргнути процесу екстракције и апстракције. Због ограничења рачунарских ресурса креиран је и помоћни скуп података, меморијски мање захтеван, где је идеја да се сумаризује апстракт и генерише наслов рада. Циљ овог приступа је да се испрати у којој мери дужина и комплексност текстуалног садржаја утиче на перформансе модела.

3.2. Претпроцесирање података

У процесу претпроцесирања података радови су конвертовани из *.pdf* формата у *.txt* формат. За сваки рад спроведено је издвајање апстракта, који се касније користе као референтни сажеци током тренирања и тестирања модела. Издвојене су кључне речи, чија сврха је да додатно помогну у евалуацији генерисаних апстракта, анализи грешака и провери да ли је модел успео да схвати главну идеју научног рада.

Апстракт и кључне речи написане на енглеском језику избачене су из текстуалне репрезентације рада. Уклоњено је заглавље рада, наслов, аутор, област истраживања, заједно са следећим поглављима: *литература*, *библиографија*, *биографија* и *захвалност*. Радови написани на ћирилици конвертовани су у латинично писмо.

Текст је очишћен од референци и специјалних карактера за набрајање, као и од *scientific* реченица које садрже математичке симболе и операторе. Знакови интерпункције и дијакритички знакови нису уклањани из текста, како би генерисани сажетак био што веродостојнији у граматичком смислу.

Сваки рад подвргнут је процесу издвајања наслова поглавља применом анализе стилова. Испоставило се да већина наслова прати форматирање *Times New Roman, Bold* и формат *број-тачка-наслов*. За анализу форматирања *.pdf* докумената кориштена је *fitz* библиотека. Мотивација за издвајање наслова лежи у томе што нису сва поглавља подједнако релевантна за апстракт. Анализа комплетног научног рада доприноси бољем схватању идеје и суштине тог рада, али приликом генерисања апстракта фокус треба да буде на најрелевантнијим поглављима за ту сврху. Истраживања су показала да 72% апстракта чине информације енкапсулиране у уводу, закључку и табелама [9].

3.3. Фаза екстракције

Фаза екстракције за циљ има да прекопира најзначајније реченице из научног рада за генерисање апстракта. Резултат ове фазе је редукована текстуална репрезентација научног рада, која служи као улаз за финалну фазу апстракције. Ова оптимизација у смањењу броја реченица значајно смањује време тренирања модела, димензионалност проблема, као и потребу за рачунарским ресурсима.

За потребе екстракције реченица имплементиран је *TextRank* алгоритам [3] који рачуна косинусну сличност између свих парова реченица и примењује се над сваким поглављем научног рада. Генерисани су сажеци од 5, 10, 15 и 20 реченица, где је идеја да се из увода и закључка копира дупло више реченица у односу на остала поглавља. Уколико се деси да алгоритам покуша да копира више реченица него што постоји у поглављу, онда се ради распоређивање броја

[1] https://www.ftn.uns.ac.rs/ojs/index.php/zbornik

реченица на преостала поглавља. Уколико је и то немогућ случај, реченице се преузимају из најдужег поглавља.

3.4. Фаза апстракције

Задатак ове фазе је да од текстуалне репрезентације креиране током екстракције генерише финални апстракт, одн. парафразиран текст. У оквиру решења примењена је енкодер-декодер архитектура, која је послужила као полазна структура за даљи развој. Претпроцесирање података за ову фазу извршено је коришћењем *keras*-ове *Tokenizer* библиотеке. На основу идентификованих речи у тренинг скупу података, *Tokenizer* гради речник где свака реч има свој јединствени индекс – кључ је реч из текста, вредност је фреквенција речи. Генерисан речник је кориштен је за конвертовање текста у нумеричку репрезентацију и касније за декодирање, одн. враћање речи у текстуалну репрезентацију. Трениран је *custom word embedding* фитовањем *Word2Vec* модела над целим скупом података, са циљем да се обогати репрезентација речи у *embedding* слоју током процеса тренирања модела за сумаризацију текста.

Имплементирана, истренирана и тестирана су три модела за сврху генерисања сажетака научних радова:

- *vanilla sequence-to-sequence* модел [10];
- *sequence-to-sequence* модел и *attention* [10];
- *google/mt5-small* трансформер модел [11].

Vanilla sequence-to-sequence модел се састоји из енкодера и декодера и представља дуални *RNN* систем. *Sequence-to-sequence* модел са *attention* механизмом представља проширење претходно представљеног модела, где енкодер користи бидирекциони *LSTM* (*BiLSTM*). Излаз енкодера обухвата информације о контексту из оба смера, чиме се добија шире разумевање улазног текста. Декодер такође користи *LSTM* и проширен је *attention*-ом који је дефинисан функцијом *one_step_attention*. Ова функција користи тежинске коефицијенте за одабир важних делова енкодованог текста. Током сваког корака декодирања користи се *attention* механизам.

Оптимизација модела током тренирања спроведена је коришћењем *sparse_categorical_crossentropy* функције губитка. Кориштен је *RMSprop* (енгл. *Root Mean Square Propagation*) оптимизатор за минимизацију *loss* функције током тренинга. Овај одабир није нужно једини прави избор – одлука је донета на основу већ постојећих истраживања која се баве проблемом сумаризације текста.

Трансформер модел *google/mt5-small* изабран је за решавање проблема сумаризације научних радова јер подржава нестандардне језике, укључујући и српски. Јавно је доступан на *Hugging Face* платформи. *Small* верзија одабрана је због меморијских ограничења. Архитектура овог модела није у потпуности позната, с обзиром на то да детаљи имплементације нису објављени, али може се претпоставити да се заснива на сличним принципима који важе за све трансформере. Примењује *self-attention* и *masked self-attention* механизме. Модел је дотрениран на тренинг скупу података научних радова писаних на српском језику. Како је одабрани трансформер предвиђен за решавање проблема сумаризације текста, одлучено је да се за функцију губитка и оптимизатор користи оно што библиотека подразумевано нуди, а то су *cross-entropy loss* и *Adam optimizer*.

4. РЕЗУЛТАТИ И ДИСКУСИЈА

Како је *Text Rank* алгоритам ненадгледан, не постоје конкретни параметри који утичу на перформансе и успешност алгоритма. У овом раду фокус је на екстрактима који се састоје од 5 реченица - дужи екстракти испоставили су се као неподобни, јер су значајно продужавали време тренирања модела. Алгоритам је успевао да генерише кохерентне сажетке који се састоје од реченица релевантних за апстракт.

Vanilla sequence-to-sequence модел трениран је и тестиран над оба скупа података, где је фокус био на томе како број *LSTM* слојева, величина улазног *batch*-а и број епоха утичу на квалитет генерисаног сажетка. Примећено је да величина *batch*-а није значајно утицала на резултате. Емпиријски је закључено да већи број епоха током тренинг фазе доприноси томе да генерисани сажеци имају кохерентнију структуру и да су семантички и граматички коректнији. Са енкодером који садржи 1 *LSTM* слој и 2500 епоха постигнути су одлични резултати над тренинг скупом података (*rougeL (f1)* = 0.7644), док је током тестирања уочено да је модел генерисао сажетке које је научио током тренинг фазе и додељивао их научним радовима из тест скупа података. Дошло је до *overfitting*-а, али је ипак охрабрујуће што су током фазе тестирања генерисани сажеци барем били у сличном домену као циљни.

Модел са 2 *LSTM*-а слоја истрениран у 200 епоха није постигао боље резултате у односу на претходни модел. Генерисане реченице биле су неисправне и неретко се дешавало да дође до значајног понављања речи. Са повећањем броја *LSTM* слојева у енкодеру јавили су и се и проблеми меморијских ограничења, те је стога за модел чији енкодер садржи 3 *LSTM* слоја одлучено да буде трениран над скупом података за генерисање наслова на основу апстракта. Нису добијени задовољавајући резултати – контекстуално нису имали везе са циљним насловом, али су донекле били семантички и граматички коректни. Проширење *sequence-to-sequence* модела *attention*-ом није донело значајне бенефите у погледу резултата, с обзиром на то да је фаза тренирања трајала веома дуго због рачунарске комплексности овог механизма и меморијских ограничења, па је одлучено да тренирање траје свега 30 епоха.

Трансформер модел истрениран у 150 епоха над скупом података који се састоји од научних радова показао се боље у односу на све претходно анализиране моделе. Може се приметити да модел има свест о контексту и да углавном генерише једну или две реченице које затим понавља више пута. Не постоје значајне разлике између тренинг и тест резултата и стиче се утисак да генерисани сажеци садрже део кључних речи научних радова и да је генерална идеја донекле очувана (Табела 4.1). Трансформер модел је такође истрениран током 100

епоха над скупом података који се састоји од наслова радова. Добијени су резултати који имају смисла и контекстуално су повезани са улазним апстрактом и циљним насловом (Табела 4.1). Не постоје значајне разлике у резултатима између тренинг и тест скупа података.

Google/mt5-small, Epochs 150, Scientific Papers Dataset	
Циљ	u ovom radu predstavljena je rekonstukcija objekta doma vojske u subotici koji nije u funkciji dugi niz godina, a nalazi se u samom centru grada. predložene su konkretne mere transformacije, kako bi prostor postao kvalitetniji i atraktivniji, a samim tim i privlačniji za posetioce.
Излаз	u ovom radu opisan je objekat doma vojske u subotici u subotici. opisan je postupak revitalizacije objekta doma vojske u subotici u subotici. opisan je postupak revitalizacije objekta u subotici u subotici u novom sadu.
Циљ	ovaj rad obuhvata analizu postojećeg stanja kao i potrebna istorijska istraživanja sa ciljem obnove vetrenjače i njenog plasiranja u širu javnost. projektom je obuhvaćeno i rešenje enterijera.
Излаз	u ovom radu opisan je i opisana konstrukcija vetrenjače u novom sadu. opisan je i opisana konstrukcija koja je vršena na osnovu svih karakteristika i karakteristika koja se koriste za njihovu upotrebu.
Google/mt5-small, Epochs 100, Scientific Titles Dataset	
Циљ	analiza faktora uspešnosti projekata
Излаз	analiza faktora na projektima
Циљ	arhitektonska studija multifunkcionalnog objekta na klisi
Излаз	projekat multifunkcionalnog objekta na klisi u novom sadu
Циљ	grype skener za testiranje ranjivosti u kontejnerima
Излаз	primena rizika od neželjenih napada aplikacije

Табела 4.1 Преглед примера резултата трансформер модела над тест скупом података

На основу претходно приказаних резултата може се закључити да је трансформер најбољи избор за решавање овог проблема. Једино у случају коришћења овог модела постоји конзистентна веза између улаза, генерисаног излаза и циљног излаза и модел се понаша слично над тренинг и тест подацима. Слободном проценом може се видети да је трансформер генерисао донекле квалитетне апстракте. Што се наслова тиче, може се приметити да постоје генерисани наслови који не одговарају у потпуности циљном, формулација је другачија, али контекст је и даље очуван. Постигнути су бољи резултати над скупом података који се тиче генерисања наслова (Табела 4.2) – модел се боље сналази када је димензионалност проблема мања.

Dataset - model	*rouge1 (f1)*	*rouge2 (f1)*	*rougeL (f1)*
Papers - 1 LSTM	0.1434	0.0156	0.1027
Titles - 3 LSTM	0.0626	0.0037	0.0584
Titles - Attention	0.1707	0.0228	0.1275
Papers – mt5	**0.2127**	**0.0477**	**0.1688**
Titles – mt5	**0.2847**	**0.1641**	**0.2755**

Табела 4.2 *Преглед постигнутих резултата*

5. ЗАКЉУЧАК

У раду је представљен систем за аутоматску сумаризацију научних радова писаних на српском језику. Решење се састоји од два модула – фазе екстракције (*TextRank* алгоритам) и фазе апстракције (*sequence-to-sequence* модели, од којих се најбоље показао трансформер *google/mt5-small*). Модели су тренирани над два скупа података, од којих су оба креирана ручно, прикупљањем научних радова из ФТН-ове архиве. Како нумеричке вредности не осликавају у најбољој мери квалитет генерисаних сажетака, извршена је ручна евалуација. Закључено је да су апстракти и наслови семантички и граматички коректни и у вези са контекстом научног рада.

Фаза екстракције текста могла би бити унапређена тако да реченице које садрже кључне речи имају већу тежину. Постоји потреба за већим и стандардизованим скуповима података на српском језику. Претпоставља се да би са богатијим скупом података и бољим рачунарским ресурсима модели били боље истренирани. Без обзира на ограничења, може се рећи да добијени резултати представљају добру полазну основу за даља истраживања у домену сумаризације текста на српском језику.

6. ЛИТЕРАТУРА

[1] Tas, O., & Kiyani, F. (2007). A survey automatic text summarization. PressAcademia Procedia, 5(1), 205-213.

[2] Cachola, I., Lo, K., Cohan, A., & Weld, D. S. (2020). TLDR: Extreme summarization of scientific documents. arXiv preprint arXiv:2004.15011.

[3] Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language proc. (pp. 404-411).

[4] Keneshloo, Y., Shi, T., Ramakrishnan, N., & Reddy, C. K. (2019). Deep reinforcement learning for sequence-to-sequence models. IEEE transactions on neural networks and learning systems, 31(7), 2469-2489.

[5] Edmundson, H. P. (1969). New methods in automatic extracting. Journal of the ACM (JACM), 16(2), 264-285.

[6] Kosmajac, D., & Kešelj, V. (2019, March). Automatic text summarization of news articles in serbian language. In 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-6). IEEE.

[7] Lin, H., & Ng, V. (2019, July). Abstractive summarization: A survey of the state of the art. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 9815-9822).

[8] Pilault, J., Li, R., Subramanian, S., & Pal, C. (2020, November). On extractive and abstractive neural document summarization with transformer language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 9308-9319).

[9] Altmami, N. I., & Menai, M. E. B. (2022). Automatic summarization of scientific articles: A survey. Journal of King Saud University-Computer and Information Sciences, 34(4), 1011-1028.

[10] Shi, T., Keneshloo, Y., Ramakrishnan, N., & Reddy, C. K. (2021). Neural abstractive text summarization with sequence-to-sequence models. ACM Transactions on Data Science, 2(1), 1-37.

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Кратка биографија:

Наташа Ивановић рођена је 1. маја 1998. године у Сомбору. Мастер рад на Факултету техничких наука из области Електротехнике и рачунарства – Интелигентни системи одбранила је 2024. године.