Proceedings of the Language Technologies for All (LT4All) , pages 5–8
Paris, UNESCO Headquarters, 5-6 December, 2019.

Using Language Technologies to Automate the UNDP Rapid Integrated Assessment Mechanism in Serbian

Vuk Batanović *, Boško Nikolić †
* † School of Electrical Engineering, University of Belgrade
* Innovation Center, School of Electrical Engineering, University of Belgrade
Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia
vuk.batanovic@ic.etf.bg.ac.rs, nbosko@etf.bg.ac.rs

Abstract

Rapid Integrated Assessment (RIA) is a United Nations Development Programme procedure involving a comparison between a country's development policy documents and the UN-defined Sustainable Development Goals. In this paper, we present the Serbian AutoRIA system that automates this procedure in Serbian, a resource-limited yet morphologically rich language. We discuss the issues regarding the preprocessing of data for this task, and the general architecture and language-related specificities of the system. We also evaluate the performance effects of various system settings using the results of a previous, manually completed RIA procedure for Serbia.

Keywords: social good, Sustainable Development Goals, semantic search, word embeddings, word2vec

Résumé

Brza integrisana procena (engl. RIA) je procedura Programa Ujedinjenih nacija za razvoj koja podrazumeva poređenje državnih strateških dokumenata o razvoju i ciljeva održivog razvoja koje su definisale Ujedinjene nacije. U ovom radu predstavljamo srpski AutoRIA sistem koji automatizuje ovu proceduru na srpskom, jeziku sa ograničenim resursima, a razvijenom morfologijom. Razmatramo probleme koji se tiču pretprocesiranja podataka za ovaj zadatak, kao i opštu arhitekturu i jezičke specifičnosti sistema. Takođe evaluiramo efekte različitih podešavanja sistema na njegove performanse koristeći rezultate ranije, ručno sprovedene RIA procedure za Srbiju.

1. Introduction

In September 2015, the United Nations adopted the Agenda for Sustainable Development by 2030, which aims to direct development policies globally. It contains 17 Sustainable Development Goals (SDGs) divided into 169 SDG targets. The Agenda addresses various global challenges, including those related to poverty, inequality, climate, environmental risks, cooperation, and peace and justice. In order to assess a country's readiness for SDG implementation, the UN Development Programme (UNDP) created the Rapid Integrated Assessment procedure (RIA). RIA involves a manual examination of laws, plans, strategies, and other relevant documents, with the aim of determining the degree of alignment between a national development framework and the goals and targets of the 2030 SDG Agenda.

Performing a Rapid Integrated Assessment requires significant human expert labor, and is, thus, costly, in terms of both time and finance. The required human expertise includes not only a high degree of domain knowledge, but also proficiency in the language in which the relevant national documents are written, which may be prohibitive factors for minor languages.

In this paper, we present a system that automates the RIA procedure in Serbian, a morphologically rich yet resource-limited language, using natural language processing (NLP). This work was carried out within the scope of a UN Development Operations Coordination Office innovation project, led by the UN Country Team in Serbia. All the resources and the programming code in Python are publicly available at the *Serbian AutoRIA* GitHub repository[1].

The remainder of the paper is structured as follows: in Section 2, we review previous work on RIA automation, as well as related NLP work for the Serbian language. We also outline a previous, manually completed RIA in Serbian. In Section 3, we discuss the preprocessing of data in Serbian and the architecture of the Serbian AutoRIA system, and in Section 4 we evaluate its results. Section 5 contains our conclusions and some potential avenues of future research.

2. Related Work

The first and, to the best of our knowledge, the only previous effort in automating the RIA procedure was a semantic search approach by Galsurkar et al. (2018). It is centered on a semantic model that compares the meaning of every SDG target description with the meaning of every sentence/paragraph from a policy document. This model was applied to data in English, using previously conducted manual RIAs and national development plans for Bhutan, Cambodia, Liberia, Mauritius, and Namibia.

Previous work on the semantic similarity of short texts in Serbian has been limited. Furlan, Batanović, and Nikolić (2013) proposed a short-text similarity method based on using string and semantic measures with term frequency weighting, and evaluated it on the paraphrase detection task. Batanović, Cvetanović, and Nikolić (2018) presented a corpus of sentence pairs in Serbian annotated with fine-grained similarity scores, and used it to evaluate several supervised and unsupervised semantic similarity models. They found that combining term frequency weighting with a part-of-speech weighting strategy proposed in (Batanović and Bojić, 2015) yielded the best results.

A manual RIA for Serbia was conducted in 2018 by eight policy experts working for close to three months. A total of 145 potentially relevant national documents were detected, but an online document file was found for only 132 of them.

[1] https://github.com/UNDP-Serbia/SerbianAutoRIA

3. Serbian AutoRIA System

Due to the very limited scope and timeframe of the project within which this research was undertaken, we focused mostly on adapting the approach of Galsurkar et al. (2018) to the resource limitations and morphological specificities of Serbian, and verifying its viability in this setting. We will first describe the particularities of preprocessing the data in Serbian, and then the functioning of the AutoRIA system.

3.1 Data Preprocessing

We encountered three main data preprocessing issues – limitations regarding data availability and readability, script variation due to the digraphia present in Serbian, and the high morphological complexity of the language.

3.1.1 Document Collection and Text Extraction

Most of the 132 policy documents collected during the manual RIA process for Serbia were in PDF format, while some were in the form of Word or Excel files. File formats became a major concern in the preparation of data for the AutoRIA system, since NLP models operate on plain texts, and the extractability of plain text from different file formats varies greatly. A specific requirement regarding text extraction was the preservation of correct separation of distinct textual units, such as list items, table cells, headings, paragraphs, etc. This was necessary, since policy documents (e.g. action plans) often contain large tables and lists, and a semantic match for an SDG target can often be found in a single list item or table cell. Similarly, a target match can be a particular document heading or paragraph.

Automatic text extraction from PDFs proved to be quite problematic – no text extraction libraries we experimented with (e.g. *PDFMiner*, *PyPDF2*, etc.) retained the correct text formatting from the original documents. Instead, table structure was lost in the output, with table cells converted out of order, making it impossible to confidently determine the boundaries between cell contents. Paragraphs were broken into multiple text lines, often in the middle of a sentence, depending on how many lines a paragraph was visually separated into within the PDF. The same issue occurred with longer list items, headings, etc. This made it impossible to detect the boundaries between individual textual units in the extracted plain text, preventing the AutoRIA system from functioning properly, as it is based on comparing the meaning of individual textual units with the meaning of each SDG target. Moreover, though most PDF files were generated via Word to PDF conversion, some were actually collated page scans. The low quality of these scans made proper text extraction using optical character recognition infeasible. All these issues prompted us to replace the PDF files with a more suitable file format. We opted for Word files, since they are the most prominent file type after PDFs in the legal domain in Serbia, and since simple copying of their contents into plain text preserves the original text layout as much as possible. We used two main sources for procuring the replacement files:

- www.pravno-informacioni-sistem.rs – the web service of the Official Gazette of the Republic of Serbia
- www.srbija.gov.rs – the official website of the Government of the Republic of Serbia

In cases when these sources were insufficient, we searched the websites of various government ministries, NGOs, etc. However, despite our best efforts, Word file equivalents could not be found for all policy documents. In such cases, we turned to alternative file formats that are still superior to PDF in terms of text extractability, like HTML and Excel. In the end, out of the initial 132 documents, Word files were found for 110, an Excel file for one, HTML files for three, while no replacements were found for the remaining 18 PDFs. Out of those 18, we used *Adobe Acrobat Pro*'s proprietary plain text extraction module on 15, while three documents were discarded due to poor text extractability. We found *Acrobat*'s implementation to be superior to that of open source libraries, though still not perfect, especially with regard to preserving the ordering of table cells and footnotes in the extracted text.

The final policy document set therefore totaled 129 documents converted into plain text. The length of the documents varied widely, from a minimum of around a dozen pages, to a maximum of circa 1500 pages.

3.1.2 Writing Script Normalization

Serbian is a digraphic language with official use of both the Cyrillic and the Latin script, so the policy document set included documents in either script. Moreover, Latin script letters were often found in Cyrillic documents, usually due to a verbatim term from one of the European languages. To avoid a model treating the same word written in different scripts as distinct words, we transliterated all Cyrillic texts into the Latin script using the *CyrTranslit* library[2].

3.1.3 Morphological Normalization

In order to reduce the effects of the morphological complexity of Serbian on data sparsity, a morphological normalizer was required. Previous work on comparing such algorithms for Serbian on various semantic tasks (Batanović and Nikolić, 2017; Batanović, Cvetanović, and Nikolić, 2018) demonstrated that a stemmer for Croatian, a closely related language, by Ljubešić, Boras, and Kubelka (2007), is usually the best-performing option. We therefore applied this stemmer, as implemented in the *SCStemmers* package (Batanović, Nikolić, and Milosavljević, 2016), to all extracted document texts and to the official translations of SDG target definitions and indicators to Serbian.

3.2 System Functioning

Per (Galsurkar et al., 2018), the semantics of an SDG target are represented by the mean of *word2vec* embeddings (Mikolov et al., 2013) of all words in its description. Similarly, the semantic representation of a textual unit (list item, table cell, paragraph, etc.) in a policy document is the mean of the embeddings of words it consists of. Vector scaling based on word TF-IDF scores is also an option. We obtained the embeddings by training the *word2vec* model using the *gensim* library (Řehůřek and Sojka, 2010) on our corpus of 129 policy documents, as Galsurkar et al. (2018) found such specialized embeddings superior to general pre-trained ones when combined with TF-IDF scaling.

Detecting alignments between a policy document and SDG targets is done by measuring the cosine similarity between the semantic vector of each textual unit in the document and the vector of each SDG target. These similarity measurements produce a list of candidate matches for each SDG target, ranked according to their semantic similarity scores. The top-ranking candidate textual units can be given

[2] https://github.com/opendatakosovo/cyrillic-transliteration

over to human experts for closer examination, or compared to existing manual RIA findings for system evaluation.

Galsurkar et al. (2018) proposed two ways of improving this basic system setup by using previously completed manual RIAs. Firstly, they suggested basing the semantics of an SDG target not only on its description, but on a "target document" which also includes all textual matches for that target found in previously completed RIAs. Secondly, they experimented with calculating word TF-IDF scores using the pool of such target documents, instead of the policy document corpus. They found that these changes generally improved system performance on English data, so we decided to evaluate them on Serbian data as well.

4. Evaluation

We first describe the evaluation metrics and data and then present the results of the Serbian AutoRIA system.

4.1 Evaluation Setup

The evaluation metric established by Galsurkar et al. (2018) is the percentage of manually detected SDG target matches that are also chosen as candidate matches by the AutoRIA system. As the number of generated candidates per target increases, this metric will tend to converge to 100%. Since there are numerous SDG targets, the main metric we use is the average percentage of true RIA matches identified by the system, across all SDG targets.

In order to perform this evaluation on Serbian data, we relied on the previously completed, manual RIA for Serbia. The true, manually detected SDG target matches had not been extracted from the documents, so we manually copied the textual units matching the first five SDGs – those under the heading "People". We extracted a total of 342 matches. However, this data was also required to create the aforementioned SDG target documents. Since the manually completed RIA for Serbia was the only such resource in the Serbian language, it was necessary to split the matched textual units into a training set, used for target document creation, and a test set, used for system evaluation.

The data had to be divided in terms of documents, with some of them being placed in the training set, and others in the test set. We tried to maximize the uniformity of SDG target match distribution between the two sets, within the range of a standard training/test split. This, however, presented a rather difficult optimization problem, because a single document often contained matches related to several SDG targets. The matches relevant to SDGs 1–5 occurred in 42 documents. After careful consideration, 31 documents were placed in the training set and 11 in the test set, dividing the data in a 75% – 25% split. Table 1 depicts the resulting distribution of SDG target matches across the two sets. As seen in the table, the overall balance of target matches between the two sets closely follows the document balance. There is some deviation from the overall balance for some SDGs, but it is not excessive and is still within the desirable range for a training/test split.

SDG	Training set	Test set	Total
1	63 (77.78%)	18 (22.22%)	81
2	97 (78.86%)	26 (21.14%)	123
3	30 (65.22%)	16 (34.78%)	46
4	43 (70.49%)	18 (29.51%)	61
5	20 (64.52%)	11 (35.48%)	31
1 – 5	253 (73.98%)	89 (26.02%)	342

Table 1: The distribution of SDG target matches

In our evaluation, the cutoff limit for the number of candidate matches returned for each SDG target was set to 300, per (Galsurkar et al., 2018), to enable some comparability between the results on two different languages. We examined the following system settings:

- Word embedding size – we considered the values of 300, 500, and 1000. Other *word2vec* hyperparameters were set to values used by Galsurkar et al. (2018).
- Using stemming or not.
- Using TF-IDF scaling for word embeddings or not; we also considered calculating TF-IDF scores using the target documents, instead of the policy document corpus.
- Placing SDG target indicators from the 2030 SDG Agenda in their respective target documents or not.
- Placing SDG target matches from the training set in their respective target documents or not.

4.2 Evaluation Results

First, we found that increasing the embedding size does not lead to performance gains, yet augments the computational cost and the time required to complete the analysis. It is thus optimal to use lower-dimensional embeddings, such as the ones with 300 dimensions, for the Serbian AutoRIA.

We then kept the embedding size and the stemming option fixed and considered the effects of other settings, and we plotted and compared the different performances, as shown in Figure 1. System designations are as follows:

- NBOW – the basic model – does not use TF-IDF scaling, SDG target indicators, nor training set target matches.
- TFIDF – a model in which TF-IDF word scaling is used, but SDG target indicators and training set target matches are not. TF-IDF values are calculated using the entire national policy document corpus.
- TFIDF + Ind – the same model as TFIDF, except it uses SDG target indicators.
- TFIDF + TS – the same model as TFIDF, except it uses SDG target matches from the training set.
- Target TFIDF + TS – the same model as TFIDF + TS, but TF-IDF values are based on target documents only.

As seen in Figure 1, TF-IDF scaling generally improves system performance. Placing SDG target indicators into the target documents, however, proves highly detrimental. The probable cause of this effect is that the indicators do not describe the *desired* state of the world (which is the focus of SDG targets), but rather the *current* state. Hence, their use tends to mislead the model into choosing candidate matches that may be related to the semantics of a target, but are not relevant in the narrower context of RIA.

On the other hand, including training set target matches into the target documents is indeed very useful, as claimed by Galsurkar et al. (2018). However, contrary to their findings, calculating TF-IDF scores using target documents, instead of the entire policy document corpus, leads to a performance drop. This divergence between English and Serbian models is likely due to different amounts of data of both types available in each language. The Serbian policy document corpus is over twice the size of the English one, but the target document set is much larger in English, as it includes target matches from five RIAs and for all SDGs. By contrast, the Serbian set is limited to the training portion of target matches from a single RIA for only five SDGs. The quality of TF-IDF weights greatly depends on the size of the corpus used to estimate them, which is why target document-based TF-IDF performs worse on Serbian data

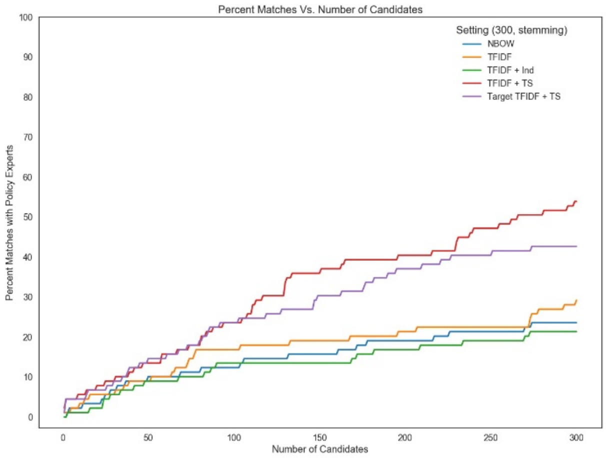

Figure 1: Effects of different settings on system results

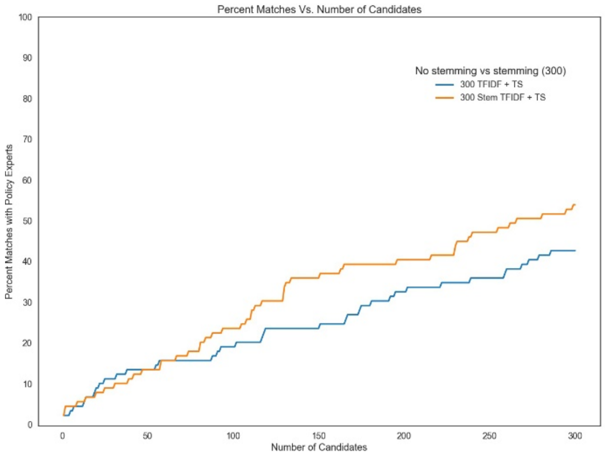

Figure 2: Effects of stemming on system results

than policy document-based TF-IDF.

In order to observe the effects of either using stemming or not, we selected the best-performing models in both settings (in both cases: TFIDF + TS) and we plotted their performances across different cutoff points. Figure 2 shows that on low cutoffs stemming does not necessarily lead to performance improvements, but as the cutoff increases, the benefits of stemming become more pronounced. At the cutoff value of 300, the model that uses stemming outperforms the one that does not by more than 10%.

While the results for the two languages are not directly comparable, this best model for Serbian seems to perform worse at lower cutoff values than the systems designed for English (Galsurkar et al., 2018). Conversely, at higher cutoffs, the system for Serbian is quite similar to the average performance of the models for English, despite being trained with much fewer previous RIA matches.

5. Conclusion

In this paper, we have presented a system for automating the Rapid Integrated Assessment procedure in Serbian, a language with limited resources yet rich morphology. We have discussed the specificities of the Serbian AutoRIA regarding data preprocessing, and the evaluation setup we used to explore the effects of different system settings.

Our findings indicate that, with careful data preprocessing and usage, promising results can be achieved even with scarce manual RIA data. Enlarging the training set would likely lead to even better performance, as would replacing the context-free word embeddings with contextual ones. As is, the current system cannot act as a substitute for human experts, but it can be a tool for assisting them in checking their findings and improving the coverage of their analysis.

6. Acknowledgements

This work was supported by the UNCT project "The use of Artificial Intelligence to automate the Rapid Integrated Assessment mechanism and to nationalize Sustainable Development Goals in Serbia". The authors closely collaborated with the SeConS Development Initiative Group and the staff of UNDP and UNCT Serbia on this project. The authors were also partially supported by the III 44009 research grant of the Ministry of Education, Science and Technological Development of the Republic of Serbia.

7. Bibliographical References

Batanović, V., and Bojić, D. (2015). Using Part-of-Speech Tags as Deep-Syntax Indicators in Determining Short-Text Semantic Similarity. *Computer Science and Information Systems*, 12(1), pp. 1–31.

Batanović, V., Cvetanović, M., and Nikolić, B. (2018). Fine-grained Semantic Textual Similarity for Serbian. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, pp. 1370–1378.

Batanović, V., and Nikolić, B. (2017). Sentiment Classification of Documents in Serbian: The Effects of Morphological Normalization and Word Embeddings. *Telfor Journal*, 9(2), pp. 104–109.

Batanović, V., Nikolić, B., and Milosavljević, M. (2016). Reliable Baselines for Sentiment Analysis in Resource-Limited Languages: The Serbian Movie Review Dataset. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, pp. 2688–2696.

Furlan, B., Batanović, V., and Nikolić, B. (2013). Semantic similarity of short texts in languages with a deficient natural language processing support. *Decision Support Systems*, 55(3), pp. 710–719.

Galsurkar, J., Singh, M., Wu, L., Vempaty, A., Sushkov, M., Iyer, D., Kapto, S., and Varshney, K. R. (2018). Assessing National Development Plans for Alignment with Sustainable Development Goals via Semantic Search. In *30th AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI 2018)*, New Orleans, Louisiana, USA, pp. 7753–7758.

Ljubešić, N., Boras, D., and Kubelka, O. (2007). Retrieving Information in Croatian: Building a Simple and Efficient Rule-Based Stemmer. In *INFuture2007: Digital Information and Heritage*, Zagreb, Croatia, pp. 313–320.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations Workshop (ICLR 2013)*, Scottsdale, Arizona, USA.

Řehůřek, R., and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, pp. 45–50.