Corpus KAS 2.0: Cleaner and with New Datasets

Aleš Žagar, Matic Kavaš, Marko Robnik-Šikonja
University of Ljubljana, Faculty of Computer and Information Science
Ljubljana, Slovenia
{ales.zagar,matic.kavas,marko.robnik}@fri.uni-lj.si

ABSTRACT

Corpus of Academic Slovene (KAS) contains Slovene BSc/BA, MSc/MA, and PhD theses from 2000 - 2018. We present a cleaner version of the corpus with added text segmentation and updated POS-tagging. The updated corpus of abstracts contains fewer artefacts. Using machine learning classifiers, we filled in missing research field information in the metadata. We used the full texts and corresponding abstracts to create several new datasets: monolingual and cross-lingual datasets for long text summarization of academic texts and a dataset of aligned sentences from abstracts in English and Slovene, suitable for machine translation. We release the corpora, datasets, and developed source code under a permissible licence.

KEYWORDS

KAS corpus, academic writing, machine translation, text summarization, CERIF classification

1 INTRODUCTION

The Corpus of Academic Slovene (KAS 1.0)[1] is a corpus of Slovenian academic writing gathered from the digital libraries of Slovenian higher education institutions via the Slovenian Open Science portal[2] [3]. It consists of diploma, master, and doctoral theses from Slovenian institutions of higher learning (mostly from the University of Ljubljana and the University of Maribor). It contains 82,308 texts with almost 1.7 billion tokens.

The KAS texts were extracted from the PDF formatted files, which are not well-suited for the acquisition of high-quality raw texts. For that reason, the KAS corpus is noisy. Our analysis showed that most original texts contain tables, images, and other kinds of figures which are transformed into gibberish when converted from the PDF format. The extracted figure captions also do not give any helpful information. Some texts contain front or back matter (for example, a table of contents at the beginning or references at the end), which shall not be present in the main text body.

The Corpus of KAS abstracts (KAS-Abs 1.0)[3] contains 47,273 only Slovene, 49,261 only English, and 11,720 abstracts in both languages. We observed several shortcomings of this corpus. A vast majority of abstracts contain keywords or the word "Abstract" somewhere in the abstract text. Many texts contain other kinds of meta-information, e.g., the name of the author or supervisor and the title of the thesis. Several corpus entries contain English and Slovene abstracts in the same unit, only one of them wrongly marked to contain both abstracts or switched Slovene and English abstracts. Several entries did not contain the abstract; instead, there was front or back matter like copyright statement, table of contents, list of abbreviations etc.

Our analysis has shown that the corpora can be improved in many aspects. Besides addressing the above-mentioned weaknesses, the main improvements in the updated KAS 2.0 and KAS-Abs 2.0 corpora are chapter segmentation and improved metadata with machine learning methods (described in Sections 2 and 3). A further motivation for our work is the opportunity to extract valuable new datasets for text summarization (monolingual and cross-lingual) and a sentence-aligned machine translation dataset created from matching Slovene and English abstracts (see Section 4). We present conclusions and ideas for further improvements in Section 5.

2 UPDATES: KAS 2.0 AND KAS-ABS 2.0

We first describe methods for extracting text and abstracts from PDF, followed by the differences between the versions 1.0 and 2.0 of corpora.

2.1 Extraction of Text Body

As many texts in corpora version 1.0 contained several hard to fix faults (like gibberish due to extracted tables and figures), we decided to extract texts once again from the PDFs. We used the pdftotext tool, which is a part of the poppler-utils. The software proved to be accurate and reliable. Its important feature is keeping the original text layout and excluding the areas where we detected figures, tables, and other graphical elements.

In the first step, we converted PDF files to images, one page at a time and used the OpenCV computer vision library to detect text and non-text areas. We marked the text areas on each page. For each document, we also calibrated the size of the header and footer areas and removed them from the text areas together with the page numbers. In this process, we removed 2,467 out of the original 91,019 documents due to the documents containing less than 15 pages or some unchecked exceptions in the code.

Next, we searched for the beginning and the end of the main text body. We observed that practically all bodies start with some variation of the Slovene word "Uvod" (i.e. introduction). If we found the beginning, we searched for the ending in the same way but with different keywords (viri, literatura, povzetek, etc). For texts with found beginning and end, the areas were clipped and the extracted texts were normalized. The normalization included handling Slovene characters with the caret (č, š, ž), ligattures (tt, ff, etc.), removal of remaining figure and table captions, and empty lines. The obtained text was segmented into the structure extracted from the table of contents. We matched headings in the text with the entries in the table of contents and used page numbers as guidelines. We ended with 83,884 successfully extracted documents.

[1] https://www.clarin.si/repository/xmlui/handle/11356/1244
[2] https://www.openscience.si/
[3] https://www.clarin.si/repository/xmlui/handle/11356/1420

Information Society 2021, 4–8 October 2021, Ljubljana, Slovenia

2.2 Extraction of Abstracts

We tried to improve the KAS-abstracts corpus by cleaning the existing documents and extracting the abstracts directly from the PDFs. An initial analysis of existing texts showed different formattings (71 different organizations publish the works in the KAS corpus). We identified five major patterns of problems and created scripts for resolving them. This produced approximately 40,000 cleaned texts while 20,000 were still problematic. The direct extraction from the PDFs followed the same procedure as for the main text body (described above). We considered figures, headers, footers, page numbers, keywords, meta-information, abstract placement at the beginning and end of the documents, multiple abstracts of different lengths, etc. This resulted in 71,567 collected Slovene abstracts. A similar procedure was applied to English abstracts and yielded 53,635 abstracts.

2.3 Differences from Version 1.0 to 2.0

Besides cleaner texts, excluded gibberish from figures and tables, and excluded front- and back-matter, the most important difference between KAS versions 1.0 and 2.0 is that the texts are segmented by structure, i.e. by headings. Unfortunately, some documents present in the original KAS were lost due to the different extraction, and for some documents appearing only in version 2.0, there is no metadata.

KAS-abstracts is greatly improved and no longer contains large quantities of unusable text and different artefacts (e.g., metadata, keywords, or front- and back-matter). Again, for some abstracts present only in version 2.0, there is no metadata. Still, they are usable for several tasks, including machine translation studies. Table 1 gives the quantitative overview of the obtained body texts and abstracts.

Table 1: Statistics of the obtained body texts and abstracts in version 2.0 of the KAS corpora.

	Sum	Same as in 1.0	Missing from 1.0	With metadata
Slo abstracts	71,567	56,610	2,383	67,533
Eng abstract	53,635	44,685	16,296	50,674
Body text	83,884	79,320	2,988	79,320

3 SUB-CERIF CLASSIFICATION

CERIF (Common European Research Information Format) is the standard that the EU recommends to member states for recording information about the research activity[4]. The top level has only five categories (humanities, social sciences, physical sciences, biomedical sciences, and technological sciences). In comparison, the lower level distinguishes 363 categories. As Slovene libraries use the UDC classification, in the KAS corpus 1.0, only 17% of the documents also contain the CERIF and sub-CERIF codes in their metadata. These are mapped from UDC codes by the heuristics produced by the Slovene Open Science Portal. Below, we describe how we automatically annotated documents with missing sub-CERIF codes using a machine learning approach.

We build a dataset for automatic annotation of sub-CERIF codes from the body texts of the documents. A document may have more than one sub-CERIF code, which means that classes are not mutually exclusive. Thus, we tackle a multi-label classification problem. In the corpus, there are 13,738 documents with high confidence levels of CERIF codes which we use in machine learning. Our dataset contains 64 labels out of 363 possible. We used 10% or 1374 samples as the test set and the remaining 90% as the training set.

As several studies have shown that recent neural embedding approaches are not yet competitive with standard text representations in document level tasks, we decided to use standard Bag-of-Words representation with TF-IDF weighting. In the preprocessing step, we lemmatized texts using CLASSLA lemmatizer[5] and removed stop-words[6] and punctuation.

We compared four classifiers. For logistic regression (LR), k-nearest neighbours (KNN), and support vector machines (SVM), we used Scikit-learn [6], and for the multi-layer perceptron (MLP), we tried Keras implementation. For the first three, we preliminary tried several different parameter values but found that they perform the best with the default ones. The MLP neural network consists of one hidden layer with 256 units, sigmoid activation function on hidden and output layers, Adam optimizer [5] with an initial learning rate of 0.01, and binary cross-entropy as a loss function. We used the early stopping (5 consecutive epochs with no improvement) and reduced the learning rate on the plateau (halving learning rate for every 2 epochs with no improvement) as callbacks during the learning process.

In Table 2, we report pattern accuracy and binary accuracy of the trained classifiers. A model predicts a correct pattern if it assigned all true sub-CERIF codes to a document. For binary accuracy, a model predicts a sub-CERIF code correctly if it assigns a true single sub-CERIF code to the document. For example, let us assume that we have four sub-CERIF codes and an example with a label sequence ’1010’. If a model predicts ’1010’, it receives 100% for both pattern and binary accuracy. If a model predicts ’0010’, it gets 0% pattern accuracy and 75% binary accuracy since it misclassified only the first label.

Table 2: Results on the sub-CERIF multi-label classification task. The best result for each metric is in bold.

Algorithm	Binary accuracy	Pattern accuracy
LR	98.48	38.36
KNN	98.52	43.75
SVM	**98.68**	**47.82**
MLP	98.66	46.58

Using the pattern accuracy metric, SVM and MLP are significantly better than KNN and LR. LR is the worst performing model, and KNN is in the middle. SVM is the best, and MLP is behind for 1.24 points. We assume that we do not have enough data for MLP to beat SVM. It is difficult to assess the models regarding binary accuracy. In the test set, we have 761 examples with 1 label, 466 with 2 labels, 107 with 3 labels, 26 with 4 labels, 10 with 5, and 4 with 6. A dummy model that predicts all zeros achieves binary accuracy of 97.51. All our models are better than this baseline, and their ranks correspond with the pattern accuracy.

We conclude that given 64 labels and 10k training instances, our best model (SVM) correctly predicts almost half of them, which is a useful result.

[4] https://www.dcc.ac.uk/resources/metadata-standards/cerif-common-european-research-information-format

[5] https://github.com/clarinsi/classla

[6] We used the list from https://github.com/stopwords-iso/stopwords-sl

4 NEW DATASETS

We created two types of new datasets, described below: summarization datasets and machine translation datasets.

4.1 Summarization Datasets

We created two new datasets appropriate for *long-text summarization* in the monolingual and cross-lingual settings. The monolingual slo2slo dataset contains 69,730 Slovene abstracts and Slovene body texts and is suitable for training Slovene summarization models for long texts. The cross-lingual slo2eng dataset contains 52,351 Slovene body texts and English abstracts. It is suitable for the cross-lingual summarization task.

4.2 Machine Translation Datasets

For the creation of a sentence-aligned *machine translation dataset*, we used the neural approach proposed by Artetxe & Schwenk [1]. The main difference to other text alignment approaches is in using margin-based scoring of candidates in contrast to a hard threshold with cosine similarity. We improved the approach by replacing the underlying neural model. Instead of BiLSTM-based LASER [2] representation, we used the transformer-based LaBSE [4] sentence representation, which has significantly improved average bitext retrieval accuracy. We used the implementation from UKPLab[7]. This approach requires a threshold that omits candidate pairs below a certain value. This value represents a trade-off between the quantity and quality of aligned pairs. The higher the threshold, the better the quality of alignments, but more samples are discarded.

In text alignment, sentences do not always exhibit one-to-one mapping: a source sentence can be split into two or more target sentences and vice versa. To address the problem, we iteratively ran the alignment process until all sentences above the chosen threshold were assigned to each other. In cases of more than one sentence assigned to a single sentence, we merged them and thus created a translation pair.

We manually inspected the alignments consisting of more than one sentence in either source or target text on a small subset of abstracts. We observed that a merging process produces better results than imposing a restriction allowing only the one-to-one mapping. In Table 4, we present an example of the alignment. The first column represents a margin-based score. If an aligned pair contains more than one sentence in the source or target, the score consists of the average margin-based score between a single sentence and multiple sentences. The last column is an indicator of whether merging was applied.

We used the ratio variant of margin-based scoring and set the default threshold to 1.1. We manually tested the alignment on our internal dataset. From 2015 examples, we successfully aligned 2002 of them (99.3%), misaligned 1 (0.1%), and omitted 12 of them (0.6%). The analysis of 12 omitted cases showed that some pairs do not match each other or are not accurate translations of each other, e.g., a large part of the original sentence is omitted, phrases are only distantly related, etc. However, approximately half of the 12 cases shall be aligned, which means that our model works very well, but conservatively and may fail for free translation pairs.

With the default value of the threshold (1.1), we produced 496.102 sentence pairs. We believe the threshold is strict enough to produce good-quality dataset (especially if compared to many other sentence alignments in existing translation datatsets). However, if one would prefer even more certain alignment, the value of the threshold can be further increased at the expense of less sentences in the datatset. We released three such datasets that reflect a trade-off between quality and quantity of the data. The sizes of the obtained datasets are available in Table 3.

Table 3: Size of the machine translation datasets based on the margin-based quality threshold.

Dataset	Threshold	Size
Normal alignment	1.1	496,102
Strict alignment	1.2	474,852
Very strict alignment	1.3	425,534

5 CONCLUSIONS

In this work, we created version 2.0 of Corpus KAS and Corpus KAS-Abstracts. We cleaned the texts and abstracts, introduced the text segmentation based on its structure, and improved the metadata. We created two new long text summarization datasets and a dataset of aligned sentences for machine translations. The latest versions of corpora and datasets are available on the CLARIN.SI. The corpora are annotated with the CLASSLA tool and released in txt, JSON and TEI formats. The source code for producing the new versions of the corpora[8] and the created datasets are publicly available[9] .

In future work, the extraction of metadata for entries where they are missing would be beneficial. There could be further improvements in cleaning the texts, and this would increase the number of available documents. When the corpora are extended with data post-2018, the software might need further modifications due to new formats and templates used in the academic works. Further experiments on the created MT datasets would clarify the setting of parameters and show if current MT systems benefit more from better quality or larger quantity of data.

ACKNOWLEDGMENTS

The research was supported by CLARIN.SI (2021 call), Slovenian Research Agency (research core funding program P6-0411), Ministry of Culture of Republic of Slovenia through project Development of Slovene in Digital Environment (RSDO), and European Union’s Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). We thank Tomaž Erjavec (JSI, Department of Knowledge Technologies) for providing data access and his assistance in building TEI format of the corpus.

REFERENCES

[1] Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3197–3203.

[7] https://github.com/UKPLab/sentence-transformers/blob/master/examples/applications/parallel-sentence-mining/bitext_mining.py

[8] https://github.com/korpus-kas

[9] KAS 2.0: https://www.clarin.si/repository/xmlui/handle/11356/1448
KAS-Abs 2.0: https://www.clarin.si/repository/xmlui/handle/11356/1449
Summarization datasets: https://www.clarin.si/repository/xmlui/handle/11356/1446
MT datasets: https://www.clarin.si/repository/xmlui/handle/11356/1447

Table 4: Examples from sentence-aligned Slovene-English abstracts.

Score	Slovene source sentence	English target sentence	Mrg
1.670	Moški pa pogosteje opravljajo opravila, ki se tičejo mehanizacije na kmetiji.	Men, however, often perform tasks related to machinery on the farm.	No
1.612	Zanimala nas je tudi prisotnost tradicionalnih vzorcev pri delu.	Additionally, I have also focused on the presence of traditional work patterns.	No
1.520	Želeli smo izvedeti, ali se kmečke ženske počutijo preobremenjene, cenjene in kako preživljajo prosti čas (če ga imajo).	I wanted to know whether rural women feel overwhelmed or valued, and how they spend their free time (if they have it).	No
1.441	Dotaknili smo se tudi problemov, s katerimi se srečujejo kmečke ženske med javnim in zasebnim življenjem.	Moreover, I have tackled the problems that rural women face when it comes to their public and private life.	No
1.437	Na koncu teoretičnega dela smo opisali še predloge za izboljšanje položaja kmečkih žensk v družbi.	At the end of the theoretical part, I have denoted further proposals for improving the situation of rural women in today's society.	No
1.388	V diplomskem delu obravnavamo položaj žensk v kmečkih gospodinjstvih v Sloveniji.	The thesis deals with the situation of women in rural households of Slovenia.	No
1.354	V empiričnem delu pa smo s pomočjo anketnega vprašalnika, na katerega so kot respondentke odgovarjale kmečke ženske, ugotavljali, kako je delo na kmetiji porazdeljeno med spoloma.	In the empirical part, I have conducted a survey on peasant women to determine the gender division of farm labour.	No
1.271	V teoretičnem delu predstavljamo pojme, kot so gospodinja, kmečko gospodinjstvo ter kmečka družina, kjer smo opisali tudi tipologijo kmečkih družin.	In the theoretical part, I have presented the following concepts: ″housewife″, ″rural household″ and ″rural family″. In addition, I have described the typology of rural families.	Yes
1.249	V nadaljevanju smo predstavili tradicionalno dojemanje kmečkih žensk, njihovo obravnavo skozi čas v slovenski literaturi, pojasnili smo procese, ki so vplivali na spremembo položaja kmečkih žensk skozi zgodovino ter se osredotočili na delo kmečkih žensk (delovni dan, delitev dela, vrednotenje dela).	I have explained the processes that have influenced the change in the situation of rural women through history and focused on their work (working day, divison of labour, work evaluation). Furthermore, I have shed light on the traditional perception of peasant women and their treatment over time in Slovene literature.	Yes
1.217	Ugotovili smo, da so tradicionalni vzorci delitve dela na kmetiji še vedno prisotni, saj smo iz analize anket in literature ugotovili, da ženske opravljajo večino del vezanih na dom in družino, to pa so gospodinjska dela in vzgoja otrok.	Hence, the majority of work related to home and family (housework and child-rearing) is performed by women. By analyzing the conducted survey and examining the literature, I have come to the conclusion that the division of farm labour more or less still follows traditional patterns.	Yes

[2] Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7, 597–610.

[3] Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2021. The KAS corpus of Slovenian academic writing. *Language Resources and Evaluation*, 55, 2, 551–583.

[4] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852*.

[5] Diederik P Kingma and Jimmy Ba. 2014. Adam: a method for stochastic optimization. In *Internationmal Conference on Representation Learning*.

[6] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12, 2825–2830.