IceSum: An Icelandic Text Summarization Corpus

Jón Friðrik Daðason
Hrafn Loftsson
Department of Computer Science

Reykjavik University
Iceland
{jond19, hrafn}@ru.is

Salome Lilja Sigurðardóttir
Þorsteinn Björnsson
Department of Icelandic and
Comparative Cultural Studies
University of Iceland
Iceland
{sls32, thb123}@hi.is

Abstract

Automatic Text Summarization (ATS) is the task of generating concise and fluent summaries from one or more documents. In this paper, we present IceSum, the first Icelandic corpus annotated with human-generated summaries. IceSum consists of 1,000 online news articles and their extractive summaries. We train and evaluate several neural network-based models on this dataset, comparing them against a selection of baseline methods. The best model obtains a ROUGE-2 recall score of 71.06, outperforming all baseline methods. Furthermore, we evaluate how the amount of training data affects the quality of the generated summaries. Our results show that while the corpus is sufficiently large to train a well-performing model, there could still be significant gains from increasing the size of the training set. We release the corpus and the models with an open license.

1 Introduction

Due to the increasing number of articles being published online every day, there is a growing need for robust Automatic Text Summarization (ATS) systems, which provide readers with a concise and fluent summary of their contents.

ATS systems are often divided into two main types (Gambhir and Gupta, 2017). First, based on the number of source documents used to generate a given summary, i.e. either *single-document* or *multi-document* summarization. In single document summarization, a single document is used for generating the summary, whereas in multi-document summarization many documents are used as the source for the generated summary.

The second type is based on the method used to generate the individual sentences in the summary, i.e. either *extractive* or *abstractive* summarization. Extractive summaries typically consist of sentence-level excerpts from the source document(s), and therefore tend to be grammatically correct and fluent. In contrast, abstractive summaries may contain words, phrases and sentences that do not occur in the original text. These summaries may also introduce grammatical errors and contain statements that are inconsistent with the source text.

Research on ATS for Icelandic has been limited to the evaluation of simple statistical methods (Christiansen, 2014) (described in Section 2). Furthermore, to our best knowledge, no ATS system is currently in use in companies or institutions in Iceland.

In this paper, we present *IceSum*, a corpus of 1,000 Icelandic news articles that can be used to train and evaluate Icelandic ATS systems. We continue previous work on summarizing Icelandic text by evaluating more recently proposed methods for extractive summarization, using neural network-based encoder-decoder models and pretrained language models. We benchmark several single-document ATS models on this dataset and compare them against previously published methods. The best performing model obtains a ROUGE-2[1] (Lin, 2004) recall score of 71.06. This is the first ATS model for Icelandic which obtains a better result than the Lead baseline method (described in Section 4), which obtains a score of 69.14.

Lemmatization is often employed as a preprocessing step for NLP tasks in Icelandic, as it dramatically reduces the size of the vocabulary. Although it has been shown to be beneficial for tasks such as named entity recognition (Ingólfsdóttir et al., 2020), information extraction (Steingrímsson et al., 2020) and machine translation (Barkarson and Steingrímsson, 2019), previous experiments with non-neural network-based models failed to show any improvement for extractive text summarization. We find that the same holds true for neural network-based models. Finally, we examine

[1] ROUGE-n refers to the overlap of n-grams between the system and the gold summaries.

Proceedings of NAACL-HLT 2021: Student Research Workshop, pages 9–14
June 6–11, 2021. ©2021 Association for Computational Linguistics

the relationship between the size of the training set and the quality of the generated summaries and find that increasing the size of the corpus would likely lead to significantly better results. We release the corpus[2] and the models[3] with an open license.

The rest of this paper is structured as follows. We discuss related work in Section 2 and present the summarization corpus in Section 3. The methods are presented in Section 4 and the experimental setup in Section 5. We present and discuss the evaluation results in Section 6, and, finally, we conclude in Section 7.

2 Related Work

A standard approach to extractive summarization involves allocating a score to each sentence, taking into account certain features, and selecting the most important sentences according to this score. Many different approaches have been proposed for this task, including statistical-based methods such as *TF-IDF* (Salton and McGill, 1986) and graph-based methods such as *TextRank* (Mihalcea and Tarau, 2004). Other methods include supervised machine learning approaches like Support Vector Machines (Hirao et al., 2002; Begum et al., 2009), Hidden Markov Models (Conroy and O'Leary, 2001), Conditional Random Fields (Shen et al., 2007), and genetic algorithms (Mendoza et al., 2014). These approaches obtain better results than purely statistical or graph-based methods (Gambhir and Gupta, 2017), but often require some feature engineering or rely on additional language resources, such as WordNet-like databases (Hirao et al., 2002), which may not be available for many low or medium-resource languages.

The use of neural network-based methods has become commonplace in ATS in recent years. One of the advantages of these methods is that the features are normally inferred automatically as opposed to being learnt with the help of hand-crafted feature templates as in feature-engineered systems. Cheng and Lapata (2016) proposed a neural network-based encoder-decoder model for extractive summarization. In their model, a Convolutional Neural Network (CNN) encoder is used to generate sentence representations which are fed to a Recurrent Neural Network (RNN) encoder that chooses which sentences to extract for the summary. This approach has been improved upon by Nallapati et al. (2017), who instead use a two-layer, bidirectional RNN, and later by Kedzie et al. (2018) who use a sequence-to-sequence model with attention. Encoder-decoder models have been shown to perform well, even with small training sets (Kedzie et al., 2018). More recently, Liu and Lapata (2019) use a pre-trained language model to generate sentence representations, and a two-layer transformer-based sequence classifier to determine which sentences should appear in the summary.

To date, there has been very limited research on text summarization for Icelandic. Christiansen (2014) evaluated the TF-IDF and TextRank (Mihalcea and Tarau, 2004) algorithms on a collection of 20 Icelandic news articles. Despite attempts at improving their performance through pre-processing (e.g., lemmatization and part-of-speech filtering), both algorithms were outperformed by a baseline summarizer which always selects the first few sentences of a document.

As presented in (Dernoncourt et al., 2018), the vast majority of existing text summarization corpora are in English. Of the 21 data sets listed in that paper, only two contain summaries in other languages than English, i.e. Arabic and Chinese. Our work, of compiling an Icelandic text summarization corpus, thus increases the pool of languages available to researchers and developers of ATS systems.

3 The Corpus

Our summarization corpus, IceSum, consists of 1,000 news articles from `mbl.is`, an Icelandic news site. This corpus is similar in size to manually annotated datasets for other languages, such as the DUC-2001 and DUC-2002 single document summarization datasets which contain 607 and 657 news texts, respectively (Kedzie et al., 2018). The goal was originally to assemble around 600 news articles, using the DUC-2001 dataset as a model. The summaries were generated by two annotators who are native speakers with a background in general linguistics and Icelandic literature. Ultimately, the total number of summarized news articles was 1,000, as mentioned above. The articles were split evenly among the two annotators, with each generating a single summary for 500 articles.

The articles in IceSum span a period of 22 years, published between 1998 and 2019, and the dataset was weighted towards more recent articles. It consists of four news categories: local (50%), world

[2] `http://hdl.handle.net/20.500.12537/96`

[3] `https://github.com/cadia-lvl/icesum`

(26%), business (14%) and sports news (10%). The summaries, generated by the two annotators, are extractive, consisting of full sentences or independent clauses from the source text. The majority of the summaries consist of full sentences, i.e. unaltered strings, ending in a full stop.

The sentences were carefully selected based on their informative value. The sentences extracted from the text more often than not contained nouns, especially proper nouns, that were of high importance for the context of the summary. If the agent of a sentence was a pronoun, the referent had to be included in earlier sentences in the summary. In this manner, the summaries always needed to be considered as a whole, rather than a series of sentences, functioning as independent entities. In the case of exceptionally long sentences, independent clauses were extracted from the sentence. In these cases, clauses were cut off right before or after a conjunction, so that the extracted clause would make an independent grammatical sentence in a summary.

The original goal was to compose summaries of 3–6 sentences for each article. Moreover, each summary was meant to contain no more than 50% of the original word count of the article itself. In the case of exceptionally long articles, the total number exceeded the original limit of 6 sentences, resulting in the upper limit of 8 sentences. This resulted in an average of 102 words per summary where the average length of the full articles was 302 words, or roughly three times longer.

4 Methods

We evaluated two types of models. First, three non-machine learning based models, which we refer to as the baseline models. Second, several neural network-based encoder-decoder models.

4.1 Baseline methods

Lead is a simple baseline method that creates a summary consisting of the first several sentences of a document. Despite its simplicity, it has historically proven to be extremely challenging for ATS models to outperform when summarizing news articles (Nenkova, 2005).

We also compared the neural network-based models against the two methods evaluated by Christiansen (2014) on Icelandic news articles, i.e. TextRank and TF-IDF. The graph-based TextRank algorithm is language-independent and requires no training. It uses co-occurrences in the text to identify similarities between sentences and uses the PageRank (Page et al., 1998) algorithm to rank each sentence. The TF-IDF algorithm assigns weights to words based on their frequency, typically obtained from a large text corpus. Sentence weights can be calculated as the average weight of the words they contain.

4.2 Encoder-decoder models

We evaluated an encoder-decoder model using four different extractors implemented in the *nnsum*[4] library:

- **Cheng & Lapata**: a unidirectional sequence-to-sequence based model where the inputs are weighed by the previous extraction probabilities (Cheng and Lapata, 2016).
- **SummaRuNNer**: a bidirectional, two-layer RNN-based sequence classifier that calculates the extraction probability based on several different sources, such as salience and position (Nallapati et al., 2017).
- **RNN**: a bidirectional, RNN-based tagging model (Kedzie et al., 2018).
- **Seq2Seq**: a bidirectional, sequence-to-sequence model with attention (Kedzie et al., 2018).

We additionally evaluate an encoder-decoder model trained using the TransformerSum[5] library, which is heavily based on the BertSum extractive text summarization model (Liu and Lapata, 2019). Sentence vectors are generated using a pre-trained language model, which is then fine-tuned with an additional classification layer.

5 Experimental Setup

We used 70% of the corpus for training, 15% for validation and 15% for testing. Each set consists of articles from the same time range and contains approximately the same proportion of news categories.

For models trained using the nnsum library, we use an averaging encoder, which obtains sentence representations by averaging out word embeddings. We used pre-trained GloVe embeddings (Pennington et al., 2014) with 300 dimensions, trained on

[4] `https://github.com/kedz/nnsum`

[5] `https://github.com/HHousen/TransformerSum`

the Icelandic Gigaword Corpus (IGC) (Steingrímsson et al., 2018), which contains approximately 1.5 billion tokens. All models are trained for 50 epochs, and we report results obtained on the test set for the model that achieved the highest ROUGE-2 recall score on the validation set.

The Transformer model was trained using IceBERT (Símonarson et al., 2021), which is fine-tuned using the TransformerSum library. We use a linear classifier and train for 5 epochs. Default settings were used for all experiments, unless otherwise noted. Like Kedzie et al. (2018), we continue adding sentences to our summary until it contains at least 100 words, and truncate summaries to 100 words when computing ROUGE scores.

We also investigated whether lemmatizing the text improves the quality of the summaries. For lemmatization, we first used ABLTagger (Steingrímsson et al., 2019) to assign part-of-speech tags to each token and then Nefnir (Ingólfsdóttir et al., 2019) to lemmatize the text. The Tokenizer[6] library was used to tokenize the source text. For models trained using nnsum, we use GloVe embeddings trained on a lemmatized version of the IGC.

The summarization methods we evaluated extract full sentences from a single document. During training, input sentences are labelled as 1 if they should be extracted and 0 otherwise. As the sentences in the gold summary contain both independent clauses and full sentences, we generated a sentence-level oracle summary for each document, using the same algorithm as Kedzie et al. (2018). For a given document, we greedily select the sentences which result in the highest possible ROUGE-2 score against the gold summary, which are then used to label the training data. We report ROUGE recall scores, calculated without stemming.

6 Results

The results of the evaluation are summarized in Table 1. The encoder-decoder model with the sequence-to-sequence extractor achieves the best performance, obtaining a ROUGE-2 score of 71.06, outperforming the Lead baseline as well as other previously evaluated methods.

For the first time, we have demonstrated an ATS system for Icelandic that outperforms baseline methods. Although Transformer-based models have obtained state-of-the-art performance for extractive summarization (Liu and Lapata, 2019; Zhong et al., 2020), the TransformerSum model does not outperform the Seq2Seq or SummaRuNNeR models in our experiments. This may be due to lack of hyperparameter tuning or the small size of the training set.

Model	**R-1**	**R-2**	**R-L**
Oracle	92.04	89.31	91.92
Lead	76.19	69.14	75.67
TextRank	60.43	47.09	59.07
TF-IDF	63.46	51.77	62.30
Cheng & Lapata	76.60	69.34	76.10
SummaRuNNer	**76.98**	**69.80**	**76.43**
RNN	76.84	69.79	76.26
Seq2Seq	**77.98**	**71.06**	**77.48**
TransformerSum	**76.80**	**69.59**	**76.23**

Table 1: ROUGE scores for all evaluated models. Scores in bold are statistically indistinguishable from the best model (paired t-test; $p < 0.05$).

As shown in Table 2, we find that lemmatizing the input text results in lower ROUGE scores. Our results are consistent with those of Christiansen (2014), who also finds that lemmatization has a negative impact on the quality of generated summaries.

Model	**R-1**	**R-2**	**R-L**
Oracle	92.04	89.31	91.92
Lead	**76.19**	**69.14**	**75.67**
TextRank	60.40	47.02	59.00
TF-IDF	62.25	50.21	61.24
Cheng & Lapata	**75.98**	**68.67**	**75.43**
SummaRuNNer	**75.82**	**68.17**	**75.20**
RNN	**76.17**	**69.07**	**75.64**
Seq2Seq	**76.33**	**69.19**	**75.80**

Table 2: ROUGE scores for all evaluated models when the text has been lemmatized. The TransformerSum model is omitted as it was pre-trained on unlemmatized text. Scores in bold are statistically indistinguishable from the best model (paired t-test; $p < 0.05$).

To estimate how the ROUGE score is affected by the size of the training set, we split it into 7 equally sized portions, each containing the same proportion of news categories. Figure 1 shows the ROUGE-2 recall score for the Seq2Seq model on the test set for a varying number of articles in the training set.

Notably, the Seq2Seq model almost matches the ROUGE-2 recall score of the Lead baseline method

[6] https://github.com/mideind/Tokenizer

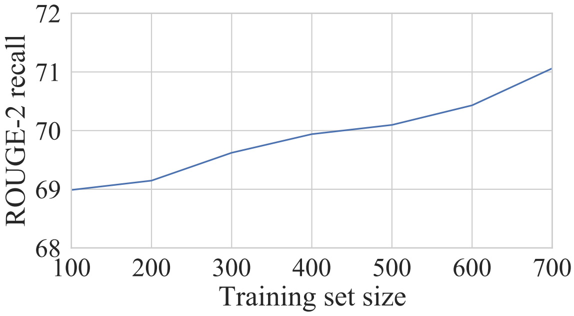

Figure 1: ROUGE-2 recall scores of the Seq2Seq model with varying amounts of training data.

with a training set of only 100 news articles. Furthermore, the ROUGE-2 score is still rising at a steady pace with a training set of 700 articles, suggesting that there may be significant benefits to enlarging the size of the corpus.

7 Conclusion

We presented the first Icelandic corpus annotated with human-generated summaries and showed that it can be used to to create an ATS system that outperforms baseline methods. We also showed that lemmatizing the source text does not result in improved performance. Finally, we evaluated how the size of the training corpus affects the quality of the generated summaries. The corpus and models have been released with an open license.

For future work, we intend to experiment further with Transformer-based models, performing hyperparameter tuning for a selection of Transformer models, such as RoBERTa-Base and ELECTRA-Base. We also plan to experiment with abstractive summarization using a much larger, unannotated corpus of Icelandic news articles. We also hope to add more summaries to the IceSum corpus in the future, and to examine inter-annotator agreement.

Acknowledgements

This work was funded by the Icelandic Strategic Research and Development Programme for Language Technology 2019, grant no. 180037-5301.

References

Starkaður Barkarson and Steinþór Steingrímsson. 2019. Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 140–145, Turku, Finland. Linköping University Electronic Press.

Nadira Begum, Mohamed A. Fattah, and Fuji Ren. 2009. Automatic Text Summarization Using Support Vector Machine. *International Journal of Innovative Computing, Information and Control*, 5(7):1987–1996.

Jianpeng Cheng and Mirella Lapata. 2016. Neural Summarization by Extracting Sentences and Words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.

Karin Christiansen. 2014. Summarization of Icelandic Texts. Master's thesis, Reykjavik University, Reykjavik, Iceland.

John M. Conroy and Dianne P. O'Leary. 2001. Text Summarization via Hidden Markov Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 406–407, New Orleans, Louisiana, USA. Association for Computing Machinery.

Franck Dernoncourt, Mohammad Ghassemi, and Walter Chang. 2018. A Repository of Corpora for Summarization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, Miyazaki, Japan. European Language Resources Association (ELRA).

Mahak Gambhir and Vishal Gupta. 2017. Recent Automatic Text Summarization Techniques: A Survey. *Artificial Intelligence Review*, 47(1):1–66.

Tsutomu Hirao, Hideki Isozaki, Eisaku Maeda, and Yuji Matsumoto. 2002. Extracting Important Sentences with Support Vector Machines. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, page 1–7, Taipei, Taiwan. Association for Computational Linguistics.

Svanhvít L. Ingólfsdóttir, Ásmundur A. Guðjónsson, and Hrafn Loftsson. 2020. Named Entity Recognition for Icelandic: Annotated Corpus and Models. In *Statistical Language and Speech Processing*, pages 46–57, Cham. Springer International Publishing.

Svanhvít Lilja Ingólfsdóttir, Hrafn Loftsson, Jón Friðrik Daðason, and Kristín Bjarnadóttir. 2019. Nefnir: A high accuracy lemmatizer for Icelandic. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 310–315, Turku, Finland. Linköping University Electronic Press.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content Selection in Deep Learning Models of Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Martha Mendoza, Susana Bonilla, Clara Noguera, Carlos Cobos, and Elizabeth León. 2014. Extractive Single-Document Summarization Based on Genetic Operators and Guided Local Search. *Expert Systems with Applications*, 41(9):4158–4169.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3075–3081, San Francisco, California, USA. AAAI Press.

Ani Nenkova. 2005. Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3*, AAAI'05, page 1436–1441, Pittsburgh, Pennsylvania. AAAI Press.

Larry Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. The PageRank citation ranking: Bringing order to the Web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2014, Doha, Qatar.

Gerard Salton and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., USA.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document Summarization Using Conditional Random Fields. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 2862–2867, Hyderabad, India. Morgan Kaufmann Publishers Inc.

Steinþór Steingrímsson, Örvar Kárason, and Hrafn Loftsson. 2019. Augmenting a BiLSTM Tagger with a Morphological Lexicon and a Lexical Category Identification Step. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP 2019, pages 1161–1168, Varna, Bulgaria. INCOMA Ltd.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, Miyazaki, Japan.

Steinþór Steingrímsson, Ágústa Þorbergsdóttir, Hjalti Daníelsson, and Gunnar Thor Örnólfsson. 2020. TermPortal: A Workbench for Automatic Term Extraction from Icelandic Texts. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 8–16, Marseille, France. European Language Resources Association.

Haukur B. Símonarson, Vésteinn Snæbjarnarson, Pétur O. Ragnarsson, and Hafsteinn Einarsson. 2021. ByteBERT: Masked language modeling for morphologically rich languages (IceBERT). Unpublished manuscript.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.