*Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 305–312
Marseille, 20-25 June 2022

# Perceived Text Quality and Readability in Extractive and Abstractive Summaries

**Julius Monsen, Evelina Rennes**
Department of Computer and Information Science
Linköping University, Linköping, Sweden
{julius.monsen, evelina.rennes}@liu.se

## Abstract

We present results from a study investigating how users perceive text quality and readability in extractive and abstractive summaries. We trained two summarisation models on Swedish news data and used these to produce summaries of articles. With the produced summaries, we conducted an online survey in which the extractive summaries were compared to the abstractive summaries in terms of fluency, adequacy and simplicity. We found statistically significant differences in perceived fluency and adequacy between abstractive and extractive summaries but no statistically significant difference in simplicity. Extractive summaries were preferred in most cases, possibly due to the types of errors the summaries tend to have.

**Keywords:** extractive summarisation, abstractive summarisation, automatic text summarisation

## 1. Introduction

Approaches to the task of automatic text summarisation can be divided into two main categories - extractive and abstractive (Hahn and Mani, 2000). In extractive summarisation, the most relevant sentences are extracted from the source document and concatenated into a summary. In abstractive summarisation, novel sentences that capture the most important information from the source document are generated.

Summarisation systems are often evaluated with content-based measures such as ROUGE and BLEU that calculate scores based on overlapping *n*-grams. These automatic measures are limited and reflect human judgements poorly (Kryściński et al., 2019). Moreover, they fail to evaluate critical features such as factual correctness, relevance of content, fluency and coherence.

Celikyilmaz et al. (2021) surveyed a range of evaluation methods for Natural Language Generation (NLG) in general. One category of methods subject to this survey was human-centric evaluation methods. Even though human evaluation is expensive to execute and the results are difficult to reproduce due to a lack of consistency in how human evaluations are run, it is deemed the most important form of evaluation for developing NLG systems. After all, the ultimate goal of NLG and text summarisation is to produce text that is valuable to people. Human evaluation is also essential as it is considered the gold standard when developing automatic measures.

There have been extensive efforts to investigate and develop new frameworks for evaluation of summaries. An example of such a framework is the FFCI framework proposed by Koto et al. (2021). Four key dimensions were identified across which to evaluate summaries. These were faithfulness (degree of factual consistency with the source), focus (precision of summary content relative to the reference), coverage (recall of summary content relative to the reference), and inter-sentential coherence (document fluency between adjacent sentences). Several traditional metrics, ROUGE included, were benchmarked through human evaluation for each dimension. The general finding was that ROUGE lacks fine-grained interpretability and that embedding-based measures correlate better with human judgement in the different dimensions.

For the above reasons, it is essential to understand what types of errors the systems make from a qualitative point of view in order to improve them. Abstractive and extractive summarisation systems suffer from different types of errors. Lux et al. (2020) showed that various factual errors are common in abstractive summaries. Extractive summarisation systems are, on the other hand, prone to produce summaries with a variety of cohesion errors (Rennes and Jönsson, 2014).

## 2. Related work

Although extractive and abstractive systems are somewhat different in how they work and what errors they produce, there have been efforts to compare certain aspects of abstractive and extractive summaries. Carenini and Cheung (2008) compared extractive and abstractive summaries focusing on the controversiality of opinions and found that the margin by which abstraction outperforms extraction is greater when controversiality is high. Moreover, Souza et al. (2021) compared extractive and abstractive summarisation methods in the task of facilitating labelling of subgroups in patent records. Nevertheless, there has been little work comparing extractive and abstractive systems regarding how users perceive text quality and readability of the summaries.

Automatic text summarisation has been proposed as one way of adapting a text to increase readability, as a shorter text could be easier to read and comprehend. For instance, Margarido et al. (2008) tested three different extraction-based summarisation strategies on target readers, and found that all strategies improved the understanding of the text to some extent. They conclude that summarisation, in combination with other techniques, could be useful for simplifying texts, but that it is important to take the literacy level of the reader into account. Smith and Jönsson (2011) showed that text complexity, given by several established text complexity measures, can be reduced by using extractive summarisation techniques. They propose summarisation as a first step to reduce the difficulty of a text, before applying other text adaptation strategies.

More recently, hybrid approaches of text simplification and summarisation have been proposed. For example, Zaman et al. (2020) adapted the Pointer generator model, a combination of abstractive and extractive summarisation models, to include a simplification factor to the loss function based on lexical complexity, and used simplified summaries as training data.

In this paper, we apply a human-centric evaluation perspective and compare aspects of perceived readability and text quality in extractive and abstractive summaries of Swedish news texts. We intend to highlight the challenges that need to be addressed from a usability perspective.

Extractive and abstractive summaries produced by two different systems are compared in an online survey. We look at how users assess the text quality and readability in summaries of news articles and investigate the strengths and weaknesses concerning perceived quality. More specifically, questions that relate to the notions of fluency, adequacy and simplicity (Wubben et al., 2012) are asked. *Fluency* is defined as the extent to which a summary contains proper grammatical sentences. *Adequacy* is defined as the extent to which a summary conveys the same meaning as the source document. Finally, *simplicity* is defined as the extent to which a summary is easy to understand.

## 3. Procedure

In this section, we describe the procedure of the study, including a description of the data, the summarisation systems used, how the survey was conducted, and how the results were analysed.

### 3.1. Data

The data used for training and evaluating the two summarisation models consisted of news articles published in *Dagens Nyheter (DN)*, Sweden's largest morning newspaper, during the years 2000–2020. Associated with each article was a preamble, here used as the summary.

Originally, the DN dataset comprises $1,963,576$ article-summary pairs, but many of these are insufficient in terms of quality for the purposes of summarisation (Monsen and Jönsson, 2021). For example, articles and summaries can be too short, or the summary may contain important contextual information that is not mentioned again in the article.

Therefore, filtering techniques as proposed by Monsen and Jönsson (2021) were applied to the dataset to build a Swedish corpus similar to the widely used English CNN/Daily Mail corpus (Nallapati et al., 2016; Hermann et al., 2015) in terms of characteristics. Article-summary pairs with articles shorter than 50 words or summaries shorter than ten words were removed, as well as article-summary pairs with a compression ratio or uni-gram novelty above 0.4 or semantic similarity below 0.4. Compression ratio is defined as the length of the summary divided by the length of the article, uni-gram novelty as the percentage of words in the summary that do not occur in the article, and semantic similarity as the cosine embedding similarity between the article and the summary computed with Sentence-BERT (Reimers and Gurevych, 2019).

This filtering yielded $349,935$ article-summary pairs that were later used for training and testing the models and producing summaries for the survey. $9,000$ of the article-summary pairs were set aside for testing and $1,000$ for validation. The average article length in the training set was 476 words/30.3 sentences, and the average summary length was 33 words/2.5 sentences.

Although this dataset has not been used to a large extent in previous studies, it is a Swedish news corpus that has properties similar to the widely used CNN/Daily Mail corpus for English and serves the purpose of doing human evaluation on abstractive and extractive summaries in Swedish, which was the aim of this study.

### 3.2. Summarisation models

The abstractive model used in this study was trained based on the methodology proposed by Rothe et al. (2020), utilising a pre-trained Swedish BERT model (Malmsten et al., 2020) to warm-start an encoder-decoder model. Both the encoder and the decoder were warm-started with the Swedish BERT model weights, which were also shared between the two components. This methodology has been shown to produce state-of-the-art results at the same time as having relatively low training costs and flexibility concerning different languages.

The warm-started model was fine-tuned in Google Colab on a Tesla V100-SXM2-16GB GPU for $300,000$ steps with a batch size of 10. The model achieved a ROUGE-1 score of 33.73 and a ROUGE-2 score of 13.31 on the test set. Compared to the evaluation results on the English CNN/Daily Mail dataset (ROUGE-1: 39.09 and ROUGE-2: 18.10), these scores are slightly lower, which most likely has to do with the fact that the summaries were preambles of news articles.

The extractive model was trained using the same pre-trained Swedish BERT model as a base. The TransformerSum[1] framework was then used to fine-tune the model on the task of extractive summarisation. The

[1] https://github.com/HHousen/TransformerSum

reason for using this framework was that it approximately corresponded to the abstractive method used with respect to performance. To use the DN dataset, which is intrinsically abstractive, it was transformed into an extractive dataset by determining the best extractive summary for each article-summary pair that maximised ROUGE scores. This was done using a Swedish tokeniser, unlike in the original implementation.

The model was subsequently fine-tuned. We used a batch size of 16 and fine-tuned the model for three epochs in Google Colab on the Tesla V100-SXM2-16GB GPU. This extractive model achieved a ROUGE-1 score of 30.83 and a ROUGE-2 score of 10.40 when extracting the top three candidate sentences.

## 3.3. Survey

The summaries were evaluated in an online survey. For the survey, 15 articles with their respective abstractive and extractive summaries were used. These were selected from the test set containing 9000 articles. One criterion for choosing articles was that the article should be between 300 and 350 words. This was considered as a reasonable article length for the purpose of the survey.

Furthermore, the aim was to have equally long summaries as well. We, therefore, chose articles with extractive summaries between 90 and 110 words, all containing four sentences. The rationale behind this length (about a third of the article) was that the summary should capture all essential information without being too concise. The lengths of the abstractive summaries were adjusted after this when being generated.

Articles in the test set were filtered on these criteria, and 15 articles were randomly sampled. Each article was read through to ensure that it was suitable and that it could be considered as a coherent text without its preamble. If an article was hard to understand due to lost context from the preamble, a new article was randomly sampled. This re-sampling was done for about half of the original articles until all articles conformed to the requirements.

Furthermore, ROUGE scores for these 15 articles and their respective summaries were calculated. The abstractive summaries had a ROUGE-1 score of 27.03, and a ROUGE-2 score of 10.77 and the extractive summaries had a ROUGE-1 score of 24.65 and a ROUGE-2 score of 7.79. As can be seen, these scores are lower than the scores on the test set. This is because the summaries generated for the survey were longer than the summaries generated for evaluation on the test set, which corresponded more closely to the summaries in the test set. However, longer summaries were deemed more appropriate for the survey since human evaluation of these would highlight the differences between respective summarisation method more clearly. This might have had some implications, as further discussed in Section 4.2.

The appendix shows examples of summaries produced by respective models for one of the used articles.

Once having 15 articles, they were divided into 5 different survey versions with 3 articles in each. The survey was distributed and shared on Facebook from the researchers´ personal accounts, resulting in a convenience sample of 37 participants. 28 of the participants had higher education for at least three years, four had postgraduate education, two higher education for less than three years, and one had elementary school level education. The answers were approximately evenly distributed between the different articles. The least answered article got four answers, while the most answered got nine. On average, each article got 7.4 answers.

The participants were presented with the original news article, as well as one extractive and one abstractive summary of the given article. They were then asked to answer the following questions regarding each summary:

(a) *The summary contains grammatically correct sentences*

(b) *The meaning of the summary conforms to the meaning of the original text*

(c) *All the important information of the original text is contained in the summary*

(d) *The summary contains superfluous information*

(e) *The summary contains words that do not fit the context*

(f) *The summary is easy to understand*

Each question was assessed using a 5-point Likert scale ranging from *Strongly Disagree* (1) to *Strongly Agree* (5).

After reading and assessing each of the three summaries, the participants were asked which of the summaries they found to be the best (*Summary 1/Summary 2/No Difference*) and they were also asked to give an explanation of their reply in a free-text field.

## 3.4. Analysis

A statistical analysis was done on the data collected from the survey. A Wilcoxon signed-rank test was used on each category to compare the differences between extractive and abstractive summaries regarding fluency, adequacy, and simplicity.

Question (a) was intended to account for *fluency*, (b)–(e) for *adequacy*, and (f) for *simplicity*. The answers for questions (d) and (e) were reversed on the 5-point Likert scale to facilitate the analysis and so that 5 had a positive connotation like in the other questions.

Inter-rater reliability was measured with Fleiss' Kappa. This was done across all articles, for all questions, and for each article respectively. Interpretations were based on thresholds proposed by Landis and Koch (1977).

Furthermore, the free-text answers motivating the choices of the best summary were analysed by reading them through and searching for overarching themes.

## 4. Results

In Table 1 mean values and standard deviations are presented for all questions and in Table 2 test statistics for all questions are presented.

| Question | Ext M | Abs M |
|---|---|---|
| (a) | 4.27 (0.953) | 3.98 (1.070) |
| (b) | 3.72 (0.962) | 2.51 (1.242) |
| (c) | 3.15 (1.169) | 2.40 (1.154) |
| (d) | 3.98 (1.144) | 3.46 (1.242) |
| (e) | 4.26 (1.059) | 3.98 (1.191) |
| (b)–(e) | 3.78 (0.817) | 3.09 (0.862) |
| (f) | 3.55 (1.241) | 3.41 (1.232) |

Table 1: Mean values (M) and standard deviations (within parenthesis) for extractive (Ext) and abstractive (Abs) summaries on all questions. The adequacy questions are presented separately and combined.

| Question | W(111) | Cohen's d |
|---|---|---|
| (a) | 577* | 0.271 |
| (b) | 465** | 0.751 |
| (c) | 841** | 0.506 |
| (d) | 582** | 0.359 |
| (e) | 523* | 0.209 |
| (b)–(e) | 900** | 0.645 |
| (f) | 1037*** | 0.094 |

** p<0.05, ** p<0.001 ***ns*

Table 2: Test statistics for all questions between extractive and abstractive summaries. Statistics for the adequacy questions are presented both separately and combined.

In analysing the answers from the survey we found a statistically significant difference between the perceived fluency in extractive ($M = 4.27$, $SD = 0.953$) and abstractive ($M = 3.98$, $SD = 1.070$) summaries, $W(111) = 577$, $p < .05$, with a small to medium effect size ($d = 0.271$). In Figure 1, this difference is illustrated.

As can be seen in Tables 1 and 2 statistically significant differences were additionally found between extractive and abstractive summaries regarding perceived adequacy in all four questions, extractive summaries having higher adequacy. When combining these four questions by averaging the scores, the difference was also statistically significant.

The difference in adequacy between extractive and abstractive summaries is illustrated by Figure 2. The plot shows the combined adequacy measure, i.e. the average of questions (b)–(e).

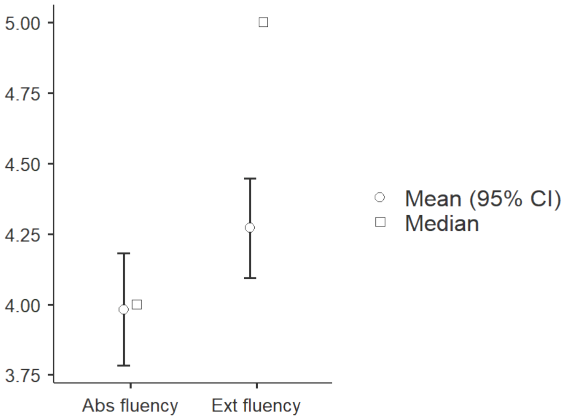

Figure 1: Difference in fluency between extractive and abstractive summaries.

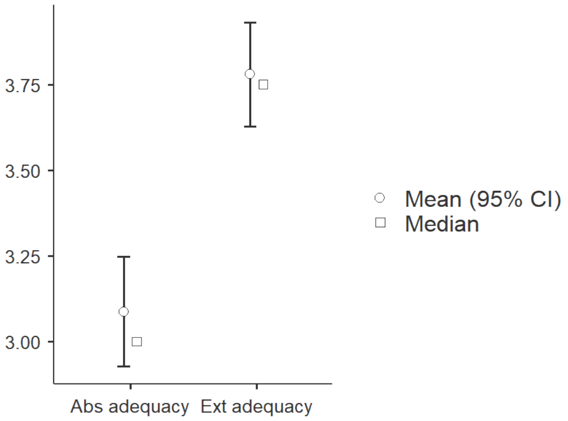

Figure 2: Difference in adequacy between extractive and abstractive summaries.

Regarding simplicity, no statistically significant difference between extractive ($M = 3.55$, $SD = 1.241$) and abstractive ($M = 3.41$, $SD = 1.232$) summaries was found, $W(111) = 1037$, $p = 0.394$ ($d = 0.094$). The difference is illustrated in Figure 3.

Furthermore, on the question of which summary the participant regarded to have better quality, the extractive summary was preferred in 74 cases, the abstractive summary in 28 cases, and in 9 cases, there was no perceived difference in terms of their quality.

Regarding inter-rater reliability, the agreement across all articles was slight ($Kappa = 0.121$, $p < 0.001$). Notably, the agreement for which type of summary was preferred was fair and substantial for abstractive and extractive summaries, respectively. For the articles separately, there was slight agreement for all cases except one for which the agreement was fair ($p < 0.05$ for all articles except for the one with only four raters).

When analysing the free-text answers, a few distinct and recurring themes stood out. First of all, the most common reasons for choosing the extractive summary as the best one were that the extractive summary was more in line with the facts presented in the article and that the abstractive summaries contained incorrect facts. The main reason for choosing the abstractive

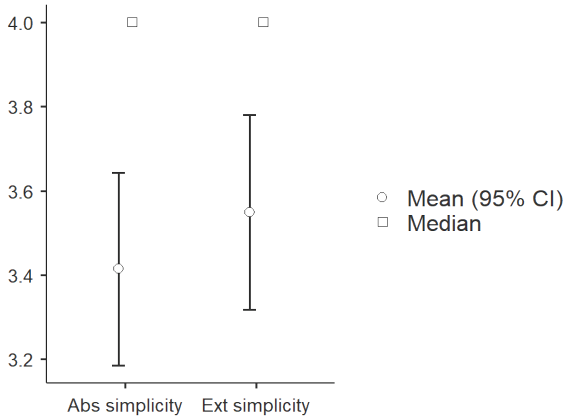

Figure 3: Difference in simplicity between extractive and abstractive summaries.

summary as the best one were that the extractive summary was poorly structured and therefore hard to follow or that it lacked some essential information.

## 5. Discussion

In this section, the results will be discussed in more detail. Certain methodological aspects are lifted as well as suggestions for future research.

### 5.1. Results

The results showed differences between the perceived text quality and readability in extractive and abstractive summaries in some regards. The statistical analysis indicated that extractive summaries were perceived as more fluent and more adequate than abstractive summaries. Although the effect size was relatively small regarding fluency, it is reasonable that the extractive summaries have higher fluency, i.e. contain more grammatically correct sentences, since the sentences are written by humans, unlike those in the abstractive summaries.

In terms of adequacy, the largest effect size was found in question (b)—whether the meaning of the summary conformed to the meaning of the original text—between extractive and abstractive summaries. This is also reasonable since extractive summaries conform to the meaning automatically as long as relevant sentences are extracted from the original text. The relatively high mean of $3.72$ ($SD = 0.962$) indicates that this was the case. The relatively low mean of $2.51$ ($SD = 1.242$) of the abstractive score furthermore indicates that abstractive summaries, to a larger extent, deviates from the meaning of the original text. This also highlights and points to the problem of factual incorrectness in abstractive summaries.

The difference in question (c)—whether all the important information of the original text was contained in the summary—was also statistically significant, with a medium effect size. Notably, this was the question where the extractive and abstractive mean values were the lowest, which may indicate that this is the aspect that needs to be improved the most in summarisation to enhance the user experience. This observation is also in line with the fact that one primary reason for choosing the abstractive summary as the best one was that vital information was missing in the extractive summary.

For question (d)—whether the summary contained superfluous information—the effect size was small to medium. The mean values of $3.98$ and $3.46$ (reversed on the 5-point Likert scale) indicate that superfluous information was not common, especially in the extractive summaries. Interestingly, this connects back to question (c), highlighting that the problem with extractive summaries does not seem to be that irrelevant sentences are extracted, but rather that relevant sentences are not extracted. This could be due to the summary length restriction. It is feasible that all relevant information of a given text can not be captured within the range of four extracted sentences.

Question (e)—whether the summary contained words that did not fit the context—also gave a significant difference between extractive and abstractive summaries, although the effect size was small. This indicates that abstractive summaries are more likely to contain words that do not fit the context. This is reasonable since the extractive summaries are indirectly written by humans. What can moreover be concluded is that the mean values ($4.26$ for extractive summaries and $3.98$ for abstractive summaries) are the highest for all questions (these values are also reversed on the 5-point Likert scale so, they are in fact, very low). This indicates that summaries rarely contained words that did not fit into the context.

For both summarisation methods, the score on simplicity, (f), is above 3, on a 5-point scale, meaning that they can be regarded as rather simple. This would mean that both summarisation methods are promising and valuable in the context of text simplification. No statistically significant difference between extractive and abstractive summaries was found regarding simplicity. Looking at Figure 3, extractive summaries show tendencies of having marginally higher simplicity on average.

The inter-rater agreement (Fleiss' Kappa) was slight in most cases. One reason for this could be that people have different backgrounds and capabilities and therefore experiences the summaries differently. However, there was a fair and substantial agreement regarding the preferred type of summary, which may indicate that the participants overall had somewhat similar views in this aspect.

These results altogether, are in line with the observation that participants preferred extractive summaries over abstractive summaries in most cases. This may have to do with the different types of errors in the summaries. As already pointed out, abstractive summaries can often be factually incorrect and deviate semantically from the original text, see for instance the abstractive summary sentence *It is now the Russian hockey league, NHL, ...* in the Appendix. In some cases, this may have been a deciding factor, which was also indicated by the free-text answers.

As was also noted, extractive summaries were sometimes perceived to be poorly structured or missing vital information. This connects to the various cohesion errors that extractive summaries are prone to produce. For example, one extractive summary used in the survey began with the sentence *"At the same time, the Iraqi government has..."*, see also the first extractive summarisation sentence in the Appendix. In other words, some context was missing at the beginning that may have been provided if the second sentence came first instead. Compared to factual errors, these errors were fewer and did not seem to matter as much in most cases.

Overall the participants rated both the extractive and the abstractive summaries relatively high on average regarding most aspects. Only two questions for abstractive summaries had an average score below 3.0. This indicates that very few summaries were considered to have low quality and that automatic text summarisation has great potential as a useful tool for end-users.

### 5.2. Methodological implications

There may have been some methodological implications. To begin with, the length of the summaries may have affected the quality of the abstractive summaries since the abstractive model was trained on shorter summaries than those used in the survey. This may have affected the model´s ability to generalise and produce longer summaries to some degree. However, as pointed out in the literature, the problems with factual incorrectness would most likely have remained. Moreover, this factor was, as illustrated, the most salient reason for choosing the extractive summary as the best one. Nevertheless, one should not rule out the possibility that the participant´s attitudes towards the summaries may have been affected by the lengths of the summaries.

It is also important to consider that this study was conducted using two specific Transformer-based models for extractive and abstractive summarisation, respectively. It is possible that other types of models would have yielded different results regarding how users perceived the summaries. Similarly, the specific dataset may also have played a role in how the models performed. For example, the preamble for a given article may not always have reflected the content in the article adequately. It is common practice to write the preambles to capture the reader's attention and interest more than inform about the content. However, since DN is considered a high-quality newspaper, this was probably mitigated to some degree.

The use of Likert scales has proven to have some limitations due to, for instance, inconsistent annotation by different annotators (see for instance Kiritchenko and Mohammad (2017)). The choice of using Likert scales for evaluation was in this case motivated by the common practice of evaluating fluency, adequacy and simplicity by such scales in automatic text simplification research.

Another essential aspect to consider is that this study did not specifically investigate how people with reading difficulties perceive summaries. Instead, a relatively homogeneous group of people, with regards to education level, participated in the study. This is important to have in mind when discussing readability and text simplification. It is hard to draw any conclusions about how useful the summaries are in terms of facilitating reading for these people since they can perceive and experience the summaries differently.

However, extractive summaries seemed to be more in line with people´s preferences based on the results of this study. Abstractive summarisation systems, as of today, still have limitations and challenges to overcome. Nevertheless, there is great potential for abstractive summarisation methods as they, in theory, can make hard-to-read sentences in a text easier to read through paraphrasing, unlike extractive summarisation methods.

### 5.3. Future research

In this study, the summary length was fixed. In future studies, it would be interesting to experiment more with the summary length to see how summaries are perceived when being longer or shorter in relation to the original text. Including participants with reading difficulties could also be very beneficial since this can lead to valuable insights into how text simplification by summarisation can be improved further. Lastly, other text genres, beyond news articles, could be explored to see if and how the attitudes and preferences differ between abstractive and extractive summaries in different settings.

## 6. Conclusion

In this study, we adopted a human-centered perspective and investigated how users perceive extractive and abstractive summaries. An online survey was conducted to compare readability and text quality aspects in summaries, namely, fluency, adequacy, and simplicity.

We found statistically significant differences between perceived fluency and adequacy in extractive and abstractive summaries, with extractive summaries having both higher fluency and adequacy. In a majority of cases, the extractive summary was further considered the best one, often with the motivation of it being more in line with the original text than the abstractive summary that often contained factual errors.

These findings highlight the need for further developing abstractive methods. Extractive methods are not perfect either, but this study has demonstrated that extraction-based summaries work well in the eyes of the user and can be a good option when it comes to text adaptation aiming to reduce text complexity.

## 7. Bibliographical References

Carenini, G. and Cheung, J. C. K. (2008). Extractive vs. NLG-based abstractive summarization of evaluative text: The effect of corpus controversiality. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 33–41, Salt Fork, Ohio, USA, June. Association for Computational Linguistics.

Celikyilmaz, A., Clark, E., and Gao, J. (2021). Evaluation of text generation: A survey.

Hahn, U. and Mani, I. (2000). The Challenges of Automatic Summarization. *Computer*, 33(11):29–36.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, Cambridge, MA, USA. MIT Press.

Kiritchenko, S. and Mohammad, S. (2017). Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, page 465–470, Vancouver, Canada.

Koto, F., Baldwin, T., and Lau, J. H. (2021). Ffci: A framework for interpretable automatic evaluation of summarization.

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., and Socher, R. (2019). Neural text summarization: A critical evaluation.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Lux, K.-M., Sappelli, M., and Larson, M. (2020). Truth or error? towards systematic analysis of factual errors in abstractive summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 1–10, Online, November. Association for Computational Linguistics.

Malmsten, M., Börjeson, L., and Haffenden, C. (2020). Playing with words at the national library of Sweden – making a Swedish BERT.

Margarido, P. R. A., Pardo, T. A. S., Antonio, G. M., Fuentes, V. B., Aires, R., Aluísio, S. M., and Fortes, R. P. M. (2008). Automatic Summarization for Text Simplification: Evaluating Text Understanding by Poor Readers. In *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*.

Monsen, J. and Jönsson, A. (2021). A method for building non-english corpora for abstractive text summarization. In *Proceedings of CLARIN Annual Conference 2021*.

Nallapati, R., Zhou, B., dos Santos, C., Gulcehre, C., and Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Rennes, E. and Jönsson, A. (2014). The impact of cohesion errors in extraction based summaries. In *9th International Conference on Language Resources and Evaluation (LREC)*, pages 1575–1582. European Language Resources Association.

Rothe, S., Narayan, S., and Severyn, A. (2020). Leveraging pre-trained checkpoints for sequence generation tasks.

Smith, C. and Jönsson, A. (2011). Automatic Summarization As Means Of Simplifying Texts, An Evaluation For Swedish. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NoDaLiDa-2010), Riga, Latvia.*

Souza, C. M., Meireles, M. R. G., and Almeida, P. E. M. (2021). A comparative study of abstractive and extractive summarization techniques to label subgroups on patent dataset. *Scientometrics*, 126:135–156.

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea, July. Association for Computational Linguistics.

Zaman, F., Shardlow, M., Hassan, S.-U., Aljohani, N. R., and Nawaz, R. (2020). Htss: A novel hybrid text summarisation and simplification architecture. *Information Processing & Management*, 57(6):102351.

## Appendix

Examples of summaries. All examples are translated from Swedish.

### Original text

Together with Canada's new star Sidney Crosby, 20-year-old Alexander Ovetchkin dueled over who would be this year's rookie in the NHL. In terms of points, the tough and skilled Russian won in Capitals, the team from the US capital. In 81 games during his first season in the NHL, he scored 52 goals and fitted for 54. 106 points in a rookie season is among the absolute best a newcomer has performed in the league. He already made his debut as a 16-year-old in the Russian league and despite allusions to fat dollar bundles from across the Atlantic, he remained at home in Moscow until last season. It also made Ovetchkin an already clear elite player when he landed in Washington, where he scored goals in his debut match. Washington as a team, however, was not enough at all to get to this year's Stanley Cup playoffs. That is why Alexander Ovetchkin can now come and show himself in the Globe already tonight and then strengthen the Russian World Cup team in Riga. In any case, this is what the Russian team leadership hopes. Ovetchkin would actually have joined the Russian team that came from Helsinki on Thursday afternoon. Russia played its first match in the Hockey Games against Finland in Helsinki and won on Wednesday night. Now Ovetchkin was delayed from the US and will not land in Stockholm until Friday morning. Alexander Ovetchkin is just one of a string of young, successful Russian hockey players who are emerging right now. Together with, among others, Ilya Kovalchuk, Atlanta and Yevgeny Malkin, Magnitogorsk in the Russian league, Russian hockey has produced stars who can take the country's hockey national team back to a level we have not seen since the early 1990s. With Ovetchkin in the team, it will be even more difficult for Sweden to win today's match in the Globe and also to take home the final victory in the Euro Hockey Tour, which will be decided on Monday.

### Extractive summary

That is why Alexander Ovetchkin can now come and show himself in the Globe already tonight and then strengthen the Russian World Cup team in Riga. Alexander Ovetchkin is just one of a string of young successful Russian hockey players emerging right now. Together with Canada's new star Sidney Crosby, 20-year-old Alexander Ovetchkin dueled over who would become this year's rookie in the NHL. With Ovetchkin in the team, it will be even more difficult for Sweden to win today's match in the Globe and also to take home the final victory in the Euro Hockey Tour, which will be decided on Monday.

### Abstractive summary

Alexander Ovetchkin is ready to play in the Russian World Cup team in ice hockey. This means that he will strengthen the Russian team in tonight's match against Sweden in the Globe. It will be Ovetchkin's first match in Moscow and it will be tougher than he had imagined before leaving the Washington Capitals to play in the Russian league. It is now the Russian hockey league, NHL, is about to take up the sport at the level not seen since the early 1990s. Today, Sweden and Russia meet in the Globe in a match that can be decisive for whether Sweden will win the Euro Hockey Tour.