A method for building non-English corpora for abstractive text summarization

Julius Monsen
Computer and Information Science
Linköping University
Linköping, Sweden
julmo634@student.liu.se

Arne Jönsson
Computer and Information Science
Linköping University
Linköping, Sweden
arne.jonsson@liu.se

Abstract

We present a method for building corpora for training, and testing, abstractive text summarizers for languages other than English. The method builds on the widely used English CNN/Daily Mail corpus and the assumption that corpora for other languages can be built by filtering language-specific news corpora to have similar properties as the CNN/Daily Mail corpus. In the paper, we show how to achieve this by removing texts from the target corpus that do not adhere to the characteristics of the CNN/DaiyMail corpus. Models are trained on these filtered subsets of the corpus and compared to results from training a model on the CNN/DaiyMail corpus. The results show that the method can be used to build corpora for training abstractive text summarisers for languages other than English that have properties on par with those trained using the CNN/Daily Mail corpus.

1 Introduction

When building, and assessing, abstractive text summarizers the English CNN/Daily Mail corpus (Nallapati et al., 2016; Hermann et al., 2015) is the currently most used benchmark corpus[1]. For languages other than English the MLSUM corpus (Scialom et al., 2020) is a corpus built on the same principles as the CNN/Daily Mail corpus. From publicly available news articles that corpus was built by filtering out articles with certain properties. This gives a multilingual extension of the CNN/Daily Mail corpus for French, German, Spanish, Russian, and Turkish.

In this paper, we present a method, similar to the one used to build the MLSUM corpus, for building a corpus in Swedish for training abstractive text summarizers. The main difference is that we go further than MLSUM and apply additional filters based on semantic textual similarity and the abstractness of the summaries to even more resemble the CNN/Daily Mail corpus. To assess the method these subsets of our corpus are used to train abstractive text summarizers and the results are compared to results achieved using the CNN/Daily Mail corpus. The corpus will be freely available as a SweCLARIN resource[2].

2 The corpora

The basis for our method is two news articles corpora, one from the Swedish newspaper Dagens Nyheter (DN) and the CNN/Daily Mail corpus. The original DN corpus comprises $1,963,576$ news articles published during the years 2000-2020. We use the preamble as a summary. Unfortunately, the preamble is not always a good summary of an article, which is one of the main problems that the proposed method handles by filtering out those article/summary-pairs that are not useful for building abstractive summarizers. Scialom et al. (2020) filter out articles shorter than 50 words or those with summaries shorter than 10 words. As we intend to apply further

[1] See e.g. https://paperswithcode.com/sota/document-summarization-on-cnn-daily-mail

[2] https://spraakbanken.gu.se/resurser#corpora

filtering mechanisms we filter out article/summary pairs with articles less than 25 words or summaries with less than 10 words. We further filter out pairs with a compression ratio (summary length/article length)[3] higher than 0.4, since very few article/summary-pairs in the CNN/Daily Mail corpus has a compression ratio higher than that.

This initial filtering and removal of duplicates yielded a corpus of 802.405 article/summary-pairs. 42% of these were categorised as domestic news, 22% as sports, 16% as economy, 10% culture and 9% other (consisting of very small categories). To further characterise the corpora we measure compression rate and novelty. Novelty (n-gram) is the fraction of n-grams in the summary that was not in the paired article (Scialom et al., 2020). In calculating novelty stop words were removed and words were stemmed. Semantic similarity was calculated as the cosine similarity between embeddings yielded from Sentence BERT (Reimers and Gurevych, 2019). Both embedding similarity on the document level (doc/doc) and between the complete summary and the most similar sentence in the article (doc/sent) are calculated. The characteristics of this corpus (DN-LC) are presented in Table 1 together with the characteristics for the CNN/Daily Mail corpus.

	DN-LC	CNN/DailyMail
Corpus size	802,405	311,971
Vocabulary size	2,575,130	803,487
Occurring 10+ times	387,370	161,820
Article words	370.41	677.21
Article sentences	24.00	28.52
Summary words	29.31	48.34
Summary sentences	2.19	3.70
Compression ratio	0.13	0.09
Novelty (uni-gram)	0.42	0.14
Novelty (bi-gram)	0.80	0.57
Novelty (tri-gram)	0.93	0.77
Semantic similarity (doc/doc)	0.49	0.65
Semantic similarity (doc/sent)	0.52	0.67

Table 1: Statistics for the different corpora. Corpus size is the number of article/summary pairs. Vocabulary size is the total number of different words in the corpus and Occurring 10+ times, the total number of words occurring 10+ times. Article/Summary words/sentences are the number of words/sentences in the articles/summaries. Compression rate, novelty, and semantic similarity are as presented in the text. All values except Corpus size, Vocabulary size and Occurring 10+ times are mean values across all article/summary-pairs.

As can be seen, the Swedish DN-LC corpus is much larger than the CNN/Daily Mail corpus. It also shows a larger variation in terms of article and summary lengths with significantly shorter articles and summaries. The Swedish corpus also has a much higher novelty. This indicates that it contains a lot of lower-quality article/summary-pairs (alternatively that the summaries are very abstract). In terms of semantic textual similarity, the Swedish corpus contains less semantically similarly article/summary-pairs. This also points to the current problem of low-quality article/summary-pairs.

3 Filtering the Swedish corpus and fine-tuning models for summarization

The goal is to create a corpus with properties similar to the CNN/Daily Mail corpus by filtering out those article/summary pairs that are not semantically similar or concrete enough, based on novelty.

[3] The reason for not measuring compression ratio as article/summary, c.f. Grusky et al. (2020) and Scialom et al. (2020) is that when filtering on compression ratio we prefer a number between 0 and 1.

We build three different corpora that are compared to the characteristics of the CNN/Daily Mail corpus. DN-S was filtered on semantic similarity first using the doc/doc similarity measure then the doc/sent similarity measure. The threshold values for filtering were iteratively adjusted to obtain a similar distribution as the CNN/Daily Mail corpus. DN-N was filtered to have similar distributions as the CNN/Daily Mail corpus with regards to novelty measures based on uni-grams, bi-grams and tri-grams. It turns, however, out that filtering on uni-grams also provides similar bi-gram and tri-gram values as well (probably since they are strongly correlated). DN-SN, finally, used both semantic similarity and novelty in a similar manner.

The results of this filtering are shown in Table 2, with the characteristics of the CNN/Daily Mail corpus as reference. The distribution among news categories was approximately the same as the DN-LC corpus for all subsets, except for DN-N which had 40% domestic news, 18% other, 17% culture, 13% economy, 12% sports.

	DN-S	DN-N	DN-SN	CNN/DailyMail
Corpus size	122,419	124,105	38,151	311,971
Vocabulary size	727,406	1,070,351	435,412	803,487
Occurring 10+ times	118,661	171,100	70,450	161,820
Article words	362.52	630.36	512.43	677.21
Article sentences	22.51	40.27	31.71	28.52
Summary words	32.15	33.16	35.67	48.34
Summary sentences	2.38	2.51	2.62	3.70
Compression ratio	0.13	0.07	0.10	0.09
Novelty (uni-gram)	0.32	0.14	0.14	0.14
Novelty (bi-gram)	0.73	0.57	0.57	0.57
Novelty (tri-gram)	0.89	0.78	0.78	0.77
Semantic similarity (doc/doc)	0.65	0.52	0.65	0.65
Semantic similarity (doc/sent)	0.67	0.60	0.67	0.67

Table 2: Statistics for the filtered corpora.

As can be seen in Table 2 the DN-SN corpus is much smaller than the CNN/Daily Mail corpus but it has the same compression ratio, novelty and semantic similarity.

4 Evaluating the corpus for abstractive summarization model building

We trained Swedish abstractive summarizers on the different corpora using a state-of-the-art approach (Rothe et al., 2020), in which an encoder-decoder model is warm-started with a pre-trained model, in our case a Swedish pre-trained BERT model (Malmsten et al., 2020). The training settings were adapted depending on the size of the corpora, but they were all trained until convergence. These models were then evaluated on two small subsets containing 9000 article/summary-pairs each, that had been picked out before the filtering. One of these subsets (test-SN) had similar semantic similarity and novelty as the CNN/Daily Mail corpus and the other (test-LC) had similar properties as the DN-LC corpus filtered on length and compression ratio. We used ROUGE scores as metrics for evaluating the model generated summaries.

In Table 3 the results for all models trained on the different corpora are presented. As can be seen, the results differ significantly between the different test sets with higher scores on test-SN for all models. Furthermore, the best result is achieved when using the rather small corpus filtered on both semantic similarity and novelty, DN-SN, with ROUGE scores almost on par with those achieved with the CNN/Daily Mail corpus. On test-SN, we also note that larger corpora, DN-S and DN-N, performs slightly worse and that the DN-LC corpus, only filtered by length and compression ratio, c.f. Scialom et al. (2020), which is even larger than the CNN/Daily Mail corpus, performs the worst. On the other hand, the model trained on the larger DN-LC corpus performs best on the test-LC set. All this highlights the importance of having high-quality test data when evaluating models and that more training data does not necessarily produce a better model, if the data is of lower quality.

	ROUGE-1		ROUGE-2		ROUGE-L	
CNN/DailyMail	39.89		18.18		27.54	
	test-LC	test-SN	test-LC	test-SN	test-LC	test-SN
DN-LC	29.08	35.46	9.67	14.71	20.23	24.71
DN-S	28.44	36.97	8.98	15.79	19.48	25.71
DN-N	27.42	36.96	8.42	16.17	18.74	25.95
DN-SN	26.48	37.22	7.44	16.32	17.84	26.11

Table 3: Evaluation results on test data measured with ROUGE F-scores.

5 Conclusion

We have presented a method for building corpora to be used when training abstractive text summarizers for Swedish. From a comparatively large corpus of Swedish summary/article pairs we use a number of filtering techniques to achieve properties similar to an English state-of-the art corpus in terms of compression ration, semantic similarity, and novelty rather than size. We show that using such a corpus to train a state-of-the art text summarizer gives results almost on par with results using the English corpus, even though the Swedish corpus is much smaller. We believe that the method can be used for other languages to build high-quality corpora that can be as useful as English corpora and extend the CLARIN infrastructure.

Acknowledgements

This research is financed by the Swedish CLARIN node SweCLARIN and the Swedish Research Council. We are indebted to Dagens Nyheter for providing the news articles.

References

Max Grusky, Mor Naaman, and Yoav Artzi. 2020. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of NAACL-HLT 2018*, pages 708–719.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 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.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of Sweden – making a Swedish BERT.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. 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.

Nils Reimers and Iryna Gurevych. 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.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online, November. Association for Computational Linguistics.