Automatic Summarization of Scientific Texts

Maria Dobko
Ukrainian Catholic University
Faculty of Applied Sciences
Lviv, Ukraine
dobko_m@ucu.edu.ua

Oleksandr Zaytsev
Ukrainian Catholic University
Faculty of Applied Sciences
Lviv, Ukraine
oleks@ucu.edu.ua

Yuriy Pryima
Ukrainian Catholic University
Faculty of Applied Sciences
Lviv, Ukraine
y.pryima@ucu.edu.ua

Abstract

In this work we overview recent advancements in the field of abstractive text summarization and discuss the use of deep learning models, especially Generative Adversarial Networks (GAN). We apply different models to our dataset of scientifix papers, acquired from arXiv, and evaluate them in terms of simplicity and performance.

We propose our own metric based on word-embeddings, which we believe will be more suitable for evaluation of abstractive summaries.

Keywords text summarization, NLP

1 Introduction

Abstractive text summarization is ... as opposed to extractive summarization where

1.1 Abstractive vs Extractive summarization

Abstractive text summarization allows us to write summaries that are almost as good as those written by humans.

Abstractive text summarization is more challenging, but it is close to what humans do.

Data-driven approach

Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning[**?**].

1.2 The power of extractive summarization

Though it will not be the focus of our work, extractive text summarization is dominant in this field and serves as a baseline for abstractive models. Before we move any further, we must understand what makes statistical techniques so powerful.

1.3 Structure of the paper

In this study we focus on abstractive text summarization. We do a brief overview of the related work that has been done in this field. Then we proceed to discussing different models for statistical and deep summarization. Then we present our dataset of scientific papers collected from arXiv and compare the performance of different models on this dataset.

2 Related work

Allahyari et al.[**?**] make a survey of the most successful text summarization techniques as of July 2017.

This September Li et al. [**?**] described their submission to the sentiment analysis sub-task of ?Build It, Break It: The Language Edition (BIBI)? where they successfully apply generative approach to the problem of sentiment analysis.

In their paper *Generative Adversarial Network for Abstractive Text Summarization* Liu et al.[**?**] built an adversarial model that achieved competitive ROUGE scores with the state-of-the-art methods on CNN/Daily Mail dataset. They compare the performance of their approach with three methods, including the abstractive model, the pointer-generator coverage networks, and the abstractive deep reinforced model.

In contrast Zhang et al.[**?**] don't use reinforcement learning but rather introduce TextGAN with an LSTM-based generator and kernelized discrepancy metric.

3 Data collection and preparation

Thanks to the open policy of arXiv we were able to collect our own dataset of scientific papers from *stat.ML* category[1]. We publish it together with this paper an encourage everyone to use it as they like in their own projects.

Structure of the data Each paper on arXiv has a unique identifier (for eample: 1801.01587). It can be used to extract any data that is available and related to that paper, including its metadata and full text as PDF.

Our dataset contains 2000 documents (papers). Each one of them is represented by a row with the following properties:

1. arXiv's unique identifier
2. title
3. abstract
4. body of the paper

Our script[2] can be used to collect more data, but for our problem a set of 2000 papers from one category is enough. For example, DUC (Document Understanding

[1] https://arxiv.org/list/stat.ML/recent

[2] Code and instructions for collecting the data is available in our repository: https://github.com/MachineLearningUCU/TextSummarization

Conference) dataset which is among the most commonly used in the feld of text summarization has 30 sets with approximately 10 documents each[3].

Cleaning the text We collected our data as PDF files and parsed them to extract the raw text. This produced a huge amount of uniterpretable UTF-8 characters from mathematical expressions. After removing those extra characters.

We wanted to remove everething that is not a known English word, but that would filter out words like "GAN", "backprop" etc. as most standard corpuses of words, such as nltk.corpus, don't include specialized scientific terminology. So we filtered the words using manually created rules based on features like word length and frequence of vowels. This leaves some noise behind, but it will not affect our models.

4 Models for text summarization

Three main classes of models that are used for text summarization task are statistical frequency computation models (TFIDF etc.), graph methods (TextRank, LexRank etc.) and machine learning approach.

4.1 Statistical methods

These methods are based on the assumption that the importance of a word or sentence in a text depends on the total number of times it appears in the document. This means that this classical approach ignores context and lexical features of the text. Furthermore, they are able to perform only extractive summarization.

4.2 Graph models

We can build a graph of each document where words or sentences are nodes and the edges are the connections between each pair of nodes. The weights on these edges represent similarity between words or sentences in the whole text. While proving to have better results than simple frequency-based methods, graph models are still bounded by the absence of lexic understanding and ability to perform extractive summary only.

4.3 Machine learning approach

In the last years, lots of attention was focused on learning how to apply neural networks to NLP tasks, including text summarization. Using encoder-decoder models it is now possible to produce abstract summaries. In [**?**] the off-the-shelf attentional encoder-decoder RNN that was originally developed for machine translation was applied to summarization and outperformed state-of-the-art systems on two different English corpora. However, there is not much information about neural networks usage in scientific text summarization.

Generative adversarial networks (GAN) are another promising approach for text summarization. Until recent years they were considered inapplicable to the discrete problems of natural language processing (NLP). The latest papers introduce novel approaches to overcoming these issues by combining GANs with reinforcement learning.

Applying generative adversarial networks to the problems of NLP is considered to be a complicated task because GANs are only defined for continuous data, and all NLP is based on discrete values like words, characters, or bytes.

However, in their latest paper Fedus, Goodfellow, and Dai[**?**] overcome this problem by using reinforcement learning to train the generator while the discriminator is still trained via maximum likelihood and stochastic gradient descent, and use it to fill the gaps in the text.

5 Evaluation metrics

Evaluating a summary is a difficult task because there is no such thing as a single summary that would be ideal for a given document. In most cases even human evaluators can not agree on which of the given summaries is better[**?**]. Unlike other NLP problems, such as translation or parsing, when it comes to text summarization we can not clearly define what makes a summary good or bad. Therefore we must make assumptions about the space of good summaries.

1. assume that a good summary would be close to some *ideal* summary manually created by humans
2. assume that the goodness of summary can be measured as the ammount of important information it contains (this assumption can be infered from the definition of text summarization).

In the following sections we briefly describe some of the commonly used metrics for text summarization that is based on the first assumption and propose our own metric that is based on second one (in fact, we will show that the proposed metric can be formulated in a different way to work with the first assumption).

5.1 ROUGE

The most widely used score for evaluating text summarizations is ROUGE (Recall-Oriented Understudy for Gisting Evaluation) inroduced by Chin-Yew Lin in 2004[**?**][**?**].

$$\text{ROUGE-N}(s) = \frac{\sum_{r \in R} \langle \Phi_n(r), \Phi_n(s) \rangle}{\sum_{r \in R} \langle \Phi_n(r), \Phi_n(r) \rangle}$$

$$\text{ROUGE-L}(s) = \frac{(1+\beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}}$$

[3] https://www-nlpir.nist.gov/projects/duc/guidelines/2001.html

ROUGE works well for extractive text summarization. But if we need to evaluate the generated summary which contains different words from the ones that occured in paper, the score will always be small because, even though the new words can be close to the expected ones, two summaries don't overlap in terms of word equality.

For example, if the human-created summary is *"The great paper"* and our model produces *"A wonderful article"*, ROUGE score will be 0, even though the summary is perfect.

5.2 Embedding-based metric

We propose a metric that uses word embeddings to evaluate summaries based on their semmantic distance to the space of good summaries. Our assumption is that a good summary of a document contains words that are semantically close to the most important words or n-grams in that document.

Let's denote an n-word summary as $S = (s_1, s_2, \ldots, s_n)$ and let $c_1, c_2, \ldots, c_k$ be the k most important n-grams in the document. For each n-gram in the generated summary we calculate the embedding-based score as its distance to the closest important n-gram in the document.

$$\text{EBM-N}(s_i) = \min_j ||s_i - c_j||_2$$

Now we define the score of the whole summary as the average score of its words

$$\text{EBM-N}(S) = \frac{1}{n} \sum_{i=1}^{n} \min_j ||s_i - c_j||_2$$

This metric is in fact the continuous version of ROUGE-N. Instead of testing the equality of n-grams in the compared summaries we use the continuous measure of semmantic distance between those n-grams.

5.2.1 Measuring word importance

To choose which words are important we can use an algorithm of extractive text summarization. We will use **term frequency-inverse document frequency** (tf-idf) as a measure of importance because of its simplicity and importance (according to [**?**], tf-idf is one of the universally used terminologies in extractive summarization).

6 Experiments and results

We have tried different methods of text summarization, both extractive (TextRank, TF-IDF) and abstractive (Seq2Seq, SeqGAN).

6.1 TextRank

In this research we used a basic model TextRank as a baseline for comparison. This is a graph method, influenced by PageRank algorithm, that represents the documents as a connected graph. We trained TextRank separately for each body of the paper and produced an extractive abstract-length summary. Scores were calculated on the basis of ROUGE score.

Table 1. Example of an abstract generated by our model

Title	Churn Prediction in Mobile Social Games: Towards a Complete Assessment Using Survival Ensembles
Real abstract	...for each player, we predict the probability of churning as function of time, which permits to distinguish various levels of loyalty profiles ... Our results show that churn prediction by survival ensembles significantly improves the accuracy and robustness of traditional analyses, like Cox regression...
Generated abstract	Conditional inference survival ensembles are constructed based on unbiased trees avoiding this problem the resulting prediction of this model contains for each player a survival function indicating the probability of churn as a function of time since the registration in the game.

6.2 Seq2Seq

We have used deep LSTM seq2seq model with attention for a task of text summarization. Seq2seq have proven to provide state-of-art result in tasks of sequence generation. At this point it is too hard to produce an abstract from the text of a paper, so we started with a simpler task of generating a title from the text of an abstract As input model takes paper abstract converted to the vectorized representation using word embeddings. The input sequence is limited by 600 words. All abstracts that are bigger than limit are omitted. All smaller abstracts are padded with ⟨SOS⟩ word that represents the end of a sequence. Model outputs sequence derived from the probability distribution. Each output word samples from this distribution having input sequence and previously generated samples. Output sequence is limited by 30 words. First ⟨SOS⟩ word represent the end of a generated summary.

7 Evaluating results

Table 2. Evaluating different models with ROUGE

Model	ROUGE-1	ROUGE-2	ROUGE-L
TextRank	18.04	4.10	10.89
LSTM	0	0	0
SeqGAN	0	0	0

Conclusions

In this paper we demonstrated how Generative Adversarial Networks can be used for abstractive text summarization.